Is gamification a suitable tool for increasing participant engagement with cognitive tests?

Jim Lumsden

A dissertation submitted to the University of Bristol in accordance with the requirements for award of the degree of Doctor of Philosophy in the Faculty of Science

School of Experimental Psychology

July 2018

Word Count: 38,053

Abstract

Over the past decade, gamification – the use of game design elements in non-game contexts – has rapidly grown in popularity, piquing the interest of researchers in many fields, including cognitive psychology. Computerised cognitive tasks are a vital data capture tool for these researchers, but participants often view these tasks as effortful and frustrating. Gamification offers a possible solution: if game elements can be incorporated into cognitive tasks without undermining their scientific validity, then data quality, intervention effects and participant retention might be improved. The purpose of this thesis was to establish whether gamification is a suitable tool for increasing participant engagement with cognitive tasks.

A systematic review revealed a literature of variable quality: findings were tentatively positive but hampered by heterogenous study designs, poor experimental controls and little attempt to methodically understand the effect of introducing game elements to a cognitive task. I therefore conducted a series of three online experiments. Each study investigated the effects of two common game design elements, points and theme, on cognitive data and participant engagement with the task.

I found that adding points to a cognitive test did not negatively impact the data collected, and improved participants' self-reported engagement with the task. In contrast, I found that graphically themed tests had a negative impact on cognitive data and did not clearly improve participant enjoyment compared to a non-game control. I found no evidence of an effect of gamification on behavioural measures of engagement (i.e., task usage); rather it seems that motivation to engage was often driven by financial incentive.

In summary, the evidence presented in this thesis suggests that gamification is not an effective method of increasing engagement with cognitive tasks. However, carefully implemented game elements can enhance participants' subjective experience while not impacting the data collected.

Authors Declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's *Regulations and Code of Practice for Research Degree Programmes* and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

Signed: .....[redacted]..... Date: ....26/07/2018.....

Conflicts of Interest

During the final year of work on this thesis, I was employed by Prolific on a part time basis. Several weeks before thesis submission, my contract became permanent, and I had begun a full-time position as Data Analyst at Prolific by the time of the PhD viva. Note that all experiments were designed, preregistered and conducted before any professional engagement with Prolific occurred.

Chapters based on Publications

Three chapters in this thesis are based on my previously published works.

Chapter 2: Lumsden J, Edwards EA, Lawrence NS, Coyle D, Munafò MR. Gamification of Cognitive Assessment and Cognitive Training: A Systematic Review of Applications and Efficacy. *JMIR Serious Games 2016 Jul 15;4(2):e11. [doi: 10.2196/games.5888]*

Chapter 3: Lumsden J, Skinner A, Woods AT, Lawrence NS, Munafò M. The effects of gamelike features and test location on cognitive test performance and participant enjoyment. *PeerJ 2016;4:e2184. [doi: 10.7717/peerj.2184]*

Chapter 5: Lumsden J, Skinner A, Coyle D, Lawrence N, Munafo M. Attrition from Web-Based Cognitive Testing: A Repeated Measures Comparison of Gamification Techniques. *J Med Internet Res 2017;19(11):e395. [doi: 10.2196/jmir.8473]*

Acknowledgements

I would like to thank my supervisors and advisors. The feedback, guidance and support I have received from Marcus Munafò and Andy Skinner has been invaluable. Not to mention the encouragement and advice of Natalia Lawrence, David Coyle, Jenny Barnett, Charlotte Housden, Kirsten Cater and Colin Dalton. This thesis would have been very different without their input.

I'd also like to thank all my colleagues in TARG and 5 Priory Road. Together they have created such a warm and loving environment, where researchers support each other, and good science and good fun go hand in hand. I'd particularly like to thank Adele Wang, Abi Mottershaw, Andy Gordon, Charlotte Buckley, David Troy, Duncan McCaig, Eleanor Kennedy, Jen Ferrar, Jane Cowen, Jasmine Khouja, Meg Fluharty, Miriam Cohen, Robyn Wootton and Stephen Hinde for being such good friends.

I'd like to thank the co-authors of my papers cited in this thesis: Chung Yen Looi, Steph Suddell, Sarah Peters, Olga Perski and Elizabeth Edwards. It has been an honour to work with them, and they have taught me so much about the importance of doing science in a team. I'd also like to acknowledge the artistic contribution of Melissa Groves, who created all the theme variant's stimuli and backgrounds in Experiments 2 and 3.

I gratefully acknowledge the joint scholarship I received from the Economic and Social Research Council (ESRC) and Cambridge Cognition. Their funding made this project possible. I'd also like to thank Prolific, who have been extremely accommodating in allowing me to finish my thesis while in their employ.

Thank you to my parents and my sister for providing such brilliant emotional and financial support. Thank you to my friends from orchestra and elsewhere, who have provided much needed distraction and ears to freak-out at. Finally, I must thank my rock, my righteousness, my adventuring companion Florence Emond. Sorry Flo, I'm still going to be obsessed with games despite this being over…

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ADHD: Attention Deficit/Hyperactivity Disorder

ANCOVA: Analysis of Covariance

ANOVA: Analysis of Variance

BF: Bayes Factor

CI: Confidence Interval

DBCI: Digital Behaviour Change Intervention

EF: Executive Function

GNG: Go/No-Go Task

HTML: Hypertext Markup Language

IMI: Intrinsic Motivation Inventory

MANOVA: Multivariate Analysis of Variance

PENS: Player Experience Needs Satisfaction scale

RT: Reaction Time

SD: Standard Deviation

SDT: Self-Determination Theory

SSD: Stop-Signal Delay

SSRT: Stop-Signal Reaction Time

SST: Stop-Signal Task

UID: User Identifier

VAS: Visual Analogue Scale

WM: Working Memory

Chapter 1: Introduction

In the 1964 Disney film Mary Poppins, the eponymous nanny convinces the unruly Banks children to tidy their room by transforming the task into a game. With a flick of her umbrella, toys spring to life and pack themselves away, beds spontaneously make themselves, and most importantly, the task becomes terrific fun.

Outside the realm of Disney, it is rather more difficult to find the fun in mundane activities. We must persist through many uninspiring tasks, unengaged and often motivated only by material reward. As a result, employers, researchers, doctors, teachers and games designers alike have devoted substantial time and money to solving this "engagement problem". The reason is simple: if we can understand how to make any activity engaging, then we will be able to deliver increased productivity, limitless research data, effective behaviour change interventions, fascinating lessons, riveting entertainment and much more.

1.1 The Engagement Problem

In recent years, health research has increasingly moved online. Ubiquitous access to the Internet has made it possible to deliver physical and mental health interventions remotely. The use of online work-marketplaces for crowdsourcing participants, such as Amazon MTurk (www.mturk.com) and Prolific (prolific.ac), combined with the growing number of platforms for delivering online cognitive assessments and questionnaires, such as Testable (www.testable.org), Gorilla (gorilla.sc), and Qualtrics (qualtrics.com), has given researchers the ability to gather data on large numbers of people in very short time spans [1–4]. These new technologies have allowed many types of health research, including psychological experiments and intervention trials, to be conducted via the web easily and inexpensively [5–7].

The issue for such web-based studies is that they must compete against the wealth of entertainment and distraction available on the Internet to attract their participants. This is made all the more difficult by the fact that dropping out of a web-based study is easier than doing so in a laboratory: a participant need only close their browser window [8,9]. Many authors have reported difficulties sustaining participant numbers for the duration of their online studies [10,11], and reviews of adherence to intervention trials have documented dropout rates of around 50% [12,13]. Whether the study in question consists of a single session or a series of sessions, the loss of participants who begin the study but do not complete it is known as attrition [13,14]. High levels of attrition may cause studies to suffer from smaller than intended sample sizes, incomplete datasets, wasted participant compensation and potentially biased results [15–18].

One explanation for these high levels of attrition is that taking part in research is often seen as tedious and dull, with any potential societal benefits being far in the future and removed from the individual participant. In the field of experimental psychology, computerised tasks are frequently used to measure cognitive performance, but participants often view them as effortful, repetitive and unengaging [19]. Cognitive tasks are also used in some types of mental health interventions, such as cognitive bias modification or executive function training. In these interventions the difficulties of cognitive tasks are compounded by the fact that many sessions of training are required to produce a result [20]. Furthermore, even if participants are willing to complete these arduous sessions of cognitive testing or training, there is evidence that a lack of participant motivation has a negative impact on the quality of data collected [21–24], and likely limits the effectiveness of interventions [25]. In short, despite the numerous research advantages offered by online cognitive tasks, we must conquer the hurdle of participant engagement in order to fully capitalise on its advantages.

1.2 Gamification

One potential solution comes from the world of video games. Millions of people spend their free time playing games every day, on computers, consoles and mobile devices [26]. Research has suggested that games are so popular because they provide easy access to a sense of engagement and self-efficacy which reality sometimes fails to deliver [27]. Although games present us with difficult challenges to overcome; they use narrative structure, complex graphics, strategic elements and intuitive rules to engross us in what could otherwise be a frustrating environment [28]. It should come as no surprise that the engaging power of games has already begun to be leveraged for purposes beyond entertainment, in the form of gamification.

Gamification has been defined as “the use of game design elements in non-game contexts” [29]: it hinges on the idea that we can use strategy, narrative, leaderboards, graphics and other game design elements to transform a mundane task into something engaging and fun. Over the past decade, gamification has been applied in a wide array of settings including the workplace [30], education [31,32], sustainability [33,34], marketing [35] and research [36] (See Sailer and colleagues [37] for a review of how gamification has been applied in a variety of fields).

Despite wide uptake, empirical evidence that gamification can actually improve engagement is currently lacking [38]. While some applications of gamification have met with success [39], other commercial attempts have been less successful [40,41]. Furthermore, heterogenous

study designs, incohesive application of theory and the wide variety of usage scenarios have complicated the interpretation and generalisation of this evidence [42,43]

Psychological researchers have embraced gamification regardless [44–46]. Initial studies have shown promise, with self-report questionnaires of participant enjoyment showing gamified cognitive tasks to be more enjoyable than their non-gamified equivalents [47–51]. Some studies have also reported that gamification can increase *behavioural* measures of engagement, such as the number of optional trials or testing blocks completed [50,52].

Furthermore, it has been suggested that gamified cognitive tasks may result in higher quality data and more effective training, simply by virtue of heightened engagement [47,53]. By using game elements that incentivise maximal performance, participants' goals might be adjusted from "completing the experiment as quickly as possible" to "succeeding at the game" [47]. The result could be data that represents the participant's true cognitive ability, rather than being confounded by low motivation.

Despite these potential advantages, there are several issues that arise specifically when gamifying the two main types of cognitive task: cognitive tests and cognitive training tasks. Cognitive tests are carefully validated measures of cognitive functioning, typically designed to assess a single cognitive construct (e.g., working memory (WM) capacity). As such, test validity is of great importance and, in many instances, the test is intimately connected with the measure it produces and the theory that defines it [54]. Even small changes to the test may cause additional cognitive load, alter participant attention, bias participants' responses, disrupt the measurement model and ultimately invalidate the test.

Cognitive training seeks to alter behaviour or improve cognition (or both) through tasks designed to induce neuronal plasticity in specific domains. Many different types of cognitive training exist [55], including cognitive bias modification [56], executive function (EF) training [57] and WM capacity training [58], and these tasks might be just as sensitive to modification as cognitive tests. Although measurement validity is not at stake, poorly thought out gamification might inadvertently shift participant attention towards superfluous game elements, thereby reducing the salience of the task elements which are driving the training effect.

In short, designing a gamified cognitive task is a balancing act. Game design elements must be introduced carefully to ensure they do not reduce training effect or threaten test validity, but they must also be substantial enough to meaningfully improve the participant experience.

1.3 Thesis Aim

This thesis aims to establish whether gamification is a suitable tool for increasing participant engagement with cognitive tests. I address this aim by answering three questions:

1. **Does the gamification of a cognitive test affect the data collected?**

   When a cognitive test is gamified, what are the effects on the primary and secondary outcome measures that it produces? Do different game elements affect the data in different ways? Could measures of participant cognition be improved through increased motivation, or does the introduction of additional cognitive load negatively affect performance?

2. **Does the gamification of a cognitive test affect participants' quality of engagement?**

   Is there any evidence that gamification improves participants' self-reported experience of the task? What are the underlying factors contributing to participant experience? Are some game elements more engaging than others?

3. **Does the gamification of a cognitive test affect participants' amount of engagement?**

   Does participants' usage behaviour change as a result of gamification? Are participants willing to test more frequently or for longer? Can gamification be used to reduce attrition from longitudinal studies?

If I can answer these questions, this thesis may provide psychologists with an evidenced method for reducing participant dropout and improving participant experience when completing cognitive tests. Regardless of gamification's effectiveness, this thesis will provide an overview of the use of gamification in the field of cognitive psychology and will shed light on the motivational influences of gamification.

1.4 Thesis Scope

The interdisciplinary nature of this thesis necessitates that I touch on research from many disparate fields. In the interests of both scope and brevity, I cannot fully examine every literature I draw upon, and therefore acknowledge several limitations of this research from the start.

I do not, for example, conduct a critical review of the theoretical basis of motivation. Nor do I perform a full review of the various theories of gamification (for reviews see [38,59–62]), or player type taxonomies (see [63–66]). These topics are beyond the scope of this work.

Although gamification has been applied in many different forms and in many different contexts (work, marketing, education, etc.). I limit my scope to cognitive psychology. Relatedly, the precise definition of gamification is still under debate [29,38], with a many different terms (gamed-up, gamelike, gameful, serious games, games with a purpose) being used, often interchangeably [47]. In the early chapters of this thesis I take a broad definition of gamification, including serious games and other purpose-built games in the literature review. In the later chapters I narrow this definition and use a stricter conception of gamification, where the game elements are relatively superficial and easily separable from the underlying task. This is the type of gamification most likely to be used by psychological researchers. A related limitation is that video games are an extremely diverse genre, with dozens of distinct game design elements and limitless combinations of said elements. Accordingly, it would be impossible for us to explore all possible gamification approaches. I instead focus my efforts on two simple game elements that are commonly used in gamified cognitive tasks: points and themes.

It is also worth noting that there is an extensive literature concerning the effect of video games on player cognition (see [67,68]). This field encompasses both the positive and negative impacts of video games on cognition, and accordingly could be considered related to the gamified cognitive training literature. I have not examined this literature, focussing instead on the application of gamification techniques to existing, validated cognitive tests.

Finally, cognitive tasks are varied in form and function. Again, it would be unrealistic to study the effect of gamification on all types of cognitive task. In Chapter 3 and beyond, I focus my research on cognitive tests designed to measure response inhibition, specifically the Go/No-Go task (GNG) and the Stop-Signal Task (SST). I consider these to be exemplar cognitive tasks: they are reaction time (RT) dependant and sufficiently difficult to be sensitive to any negative effects of gamification.

1.5 Thesis Overview

This thesis consists of seven chapters, of which this is Chapter 1. Chapter 2 systematically reviews studies that have made use of gamified cognitive tasks. I sought to understand the current state of the field, examining methods for measuring engagement and investigating the theoretical basis of gamification. I found the literature to be small and heterogenous, yet growing rapidly. Many studies reported that gamified tasks were well received by participants, and researchers were enthusiastic about gamification's potential for increasing engagement, particularly in populations that are typically difficult to engage, such as children with Attention

Deficit/Hyperactivity Disorder (ADHD). Evidence in favour of gamification was generally positive, but I saw some detrimental effects on cognitive measures and study designs were of variable quality. Only a few studies had attempted to unpick the effect of individual game design elements on engagement and cognitive data, and as such, game elements were seemingly chosen in an ad-hoc manner.

In light of these findings, I set out to investigate the impact of gamification on cognitive tests in a more rigorous way. I selected two common game elements (points and theme) and a subset of cognitive tests (assessments of response inhibition) to act as vehicles for my research. Chapter 3 documents Experiment 1: an empirical study into the effects of these game elements on cognitive data and participant enjoyment of a gamified test of response inhibition.

As my task design requirements evolved, it became apparent that existing platforms for building online cognitive tests (Gorilla.sc, Xperiment.mobi, etc) did not provide the level of customisation I desired. To address this need, I developed my own online platform for delivering gamified cognitive tasks: Mindgames. Chapter 4 documents the development of this platform and discusses the technical challenges I faced. For example, ensuring accurate stimulus presentation times, providing data security to a relatively open system, and supporting hundreds of simultaneous participants.

Chapters 5 and 6 describe Experiments 2 and 3, which both used this platform. Experiment 2 investigated whether gamification could reduce attrition from a longitudinal, online cognitive testing study, and compared behavioural and self-report measures of engagement. Experiment 3 examined whether gamification could increase the number of trials voluntarily completed by participants in a single testing session. I also explored the potentially confounding role of financial incentive in motivating participants to complete online studies.

Chapter 7 synthesises my findings and reflects on recent advances in the literature. I discuss implications for the development of gamified cognitive tests in the future, and for the field of gamification more generally. I conclude by evaluating the limitations of this thesis and suggesting future directions for research.

Chapter 2: Background and systematic review

This chapter is based on my publication in JMIR Serious Games [69].

2.1 Chapter Aims

This chapter systematically reviews how gamification has already been used for the purposes of cognitive testing and training. The goal of this review was to understand the current state of the field, investigate the theoretical basis of gamification and examine how study authors' measured engagement. I hoped to identify areas for future research, and was specifically interested in the following questions:

1. Why have researchers used gamification?
2. What cognitive domains has gamification been applied in?
3. What game design elements have been used?
4. What theory has guided gamification's application?
5. How have researchers measured engagement?
6. Based on the current evidence, does gamification work?

2.1.1 Defining Gamification

I began the review by establishing a definition of gamification; however, it quickly became apparent that the term remains subject to debate [37,38,70,71]. The most popular definition is "the use of game design elements in non-game contexts" [29], but this does not clarify how or if gamification should be delineated from the range of related terminology. For example, the tasks in this review have been variously described as 'serious games', 'gamelike', 'gamified', 'games with a purpose', 'gamed-up' or simply 'computer-based' [47,70,72,73].

I conceptualise these terms on a spectrum of 'gamelikeness'. At one end, we have serious games: fully fledged games that look and feel like commercial video games. Serious games have existed for decades with mostly educational and government origins [74–76]. Replete with game design elements (points, levels, graphics, sound, etc.), they are designed to be entertaining, but their primary purpose is to educate, train or assess behaviour [77]. For a review of the empirical evidence on the use of serious games, see [78].

At the opposite end of the spectrum, we have mundane tasks: pure in their purpose and concerned with functionality rather than entertainment. Gamification lies between these two poles: it is the process of making a mundane task more gamelike through the use of game design elements [37,79]. When we are gamifying a task, we are not so much creating a game as we are employing game elements to foster motivation [71,80].

Early efforts in gamification stayed close to the mundane end of the spectrum; using mostly superficial modifications such as points, badges and leaderboards to facilitate engagement [81,82]. However, this trio of game elements has since been heavily criticised by the game design community [83] for "taking the thing that is least essential to games and representing it as the core of the experience" [84–86]. In response, some gamification designers have pushed towards the serious games end of the spectrum: building richer game-like tasks, with systems of game elements that facilitate strategy, tactics and encourage playful thinking [84,87,88].

To ensure this review included tasks across the full breadth of the spectrum, I took a deliberately broad stance. I considered a task to be gamified if it included one or more game design elements with the purpose of increasing user motivation. I reviewed only the peer-reviewed literature and have therefore excluded gamified cognitive tasks available on the iTunes or Play Store (such as Luminosity and Peak), which are not supported by peer-reviewed research.

2.2 Methods

The following databases were searched electronically: PsycInfo, Medline, ETHOS, Embase, Pubmed, IBSS, Francis, Web of Science and Scopus. I searched the titles, abstracts and keywords of database entries using the search strategy (gamif* OR game OR games) AND (cognit* OR engag* OR behavi* OR health* OR attention OR motiv*), where * represents a wildcard to allow for alternative suffixes. Searches included articles published in English between January 2007 and October 2015. I searched the bibliographies of included articles to locate further relevant material not discovered in the database search.

2.2.1 Inclusion Criteria

- *Primary research article:* Included articles were empirical research studies, not literature reviews, opinion pieces or design documents.
- *Novel gamified task:* Included articles focussed on newly developed gamified tasks, created specifically for the study in question. I excluded commercially available video games (i.e., "off the shelf" games) as well as gamified tasks that have been in use for many years, such as Space Fortress (see [75]).
- *Measure or train cognition:* Included articles focussed on tasks designed to assess or train cognition. For scoping purposes I took a narrow definition of cognition: those processes involved in memory, attention, decision making, impulse control, executive functioning, processing speed and visual perception.

- *Validated or Piloted:* Included articles had to involve an empirical study, either validating the task as a measure of cognition or piloting the intervention. Articles regarding usability testing alone were excluded.

2.2.2 Exclusion Criteria

- *Non-peer reviewed articles:* I excluded non-peer reviewed articles such as abstracts or conference posters.
- *Gamification in the behavioural sciences but not involving cognition:* I excluded articles on gamification for education purposes, disease management, health promotion, exposure therapy or rehabilitation.
- *Game Engines/3D Environments:* I excluded articles that made use of virtual reality or a 3D environment without any game design elements or gamelike framing.

Where there was insufficient detail to determine whether an article met the inclusion criteria, I erred on the side of caution in order to increase my confidence in the relevance of the studies reviewed.

2.2.3 Data Extraction

Following screening, data were extracted from each article using a standardised data extraction form. Data relating to the application of gamification, approach taken, and efficacy was extracted from each article. *Application of gamification* refers to the field of psychology in which the gamified task was used, why a gamified task was used, and what (if any) theory guided the gamification. *Approach taken* refers to the specific game elements used in the task, and what themes and scaffolding were applied. Finally, *efficacy* refers to the findings of the study as well as details of the participants, evaluation methods, engagement measures and limitations of the study. Categorisation of concepts (such as the cognitive domains measured) was done using the article-authors' own words where possible.

All articles identified by the search strategy were screened by one reviewer (JL) in three stages, to determine whether they were relevant based on the inclusion/exclusion criteria: title, abstract and full text. A second reviewer (EE) re-screened 20% of the articles from the title stage onwards to ensure no relevant articles were missed. Articles were only included in the review on the agreement of both JL and EE.

2.3 Included Articles

My initial search yielded 33,445 articles (excluding duplicates). Of these, 23 papers from the original search and 4 papers from the manual reference search were included in the review. I repeated the search in October 2015, including articles from Jan 2015 until October 2015. This

search produced 4,448 articles (excluding duplicates) and resulted in another 4 articles being included in the review, with a further 2 also included following peer-review. The total number of review articles was therefore 33. See Figure 2.1 for a flowchart of the combined searches and Table 2.1 for details of all included studies. I used Cohen's K to assess inter-rater reliability of article inclusion at the 20% data check stage (7,590 articles checked). There was moderate agreement between the two reviewers (k=.526, 95% Confidence Interval (CI), .416 to .633, p<.001).

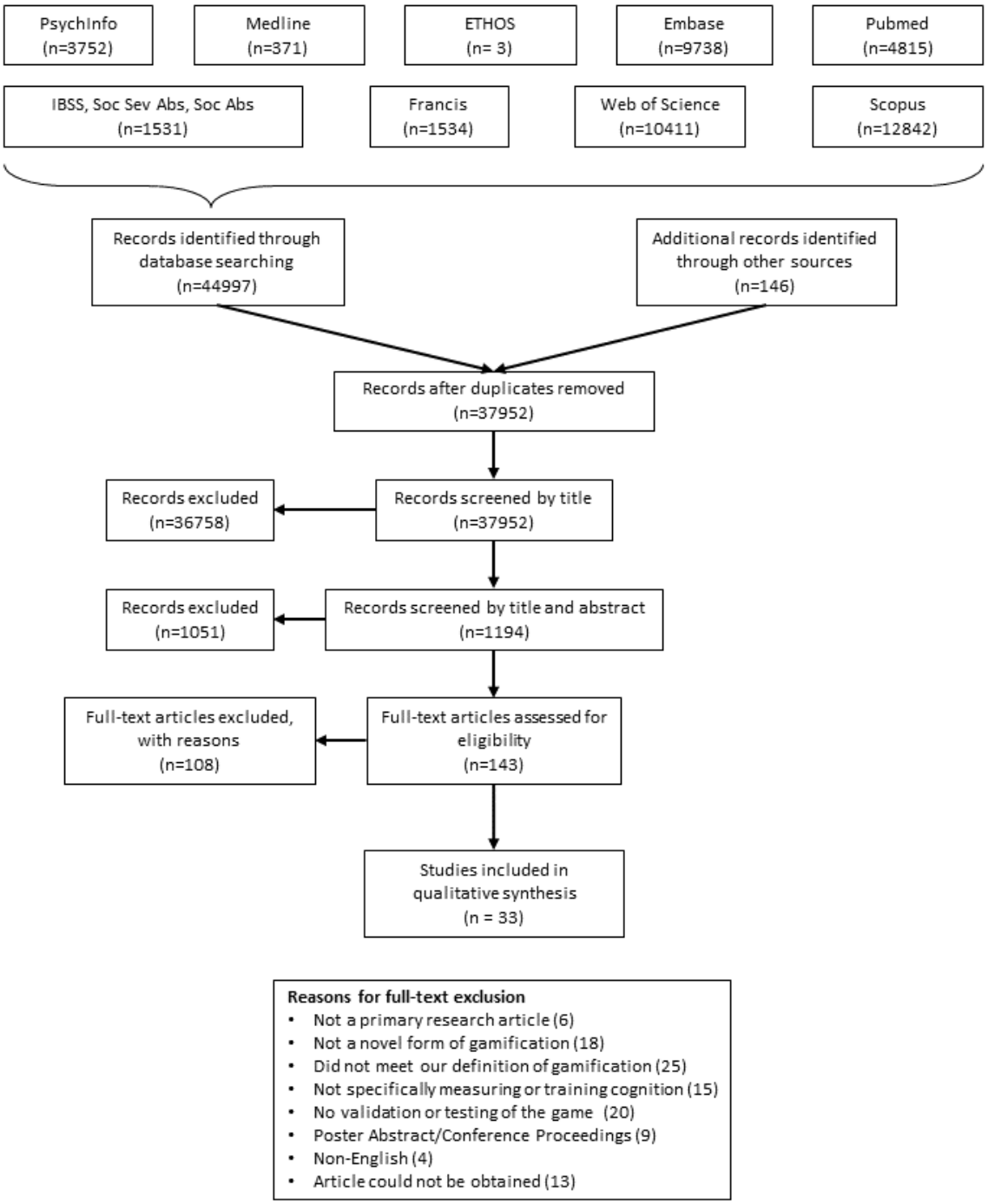


Figure 2.1 Flow chart detailing the article discovery and screening process

Table 2.1 Details of included studies. In cases where the game was not named, I assigned a descriptive name.

Author, Year	Full Title	Game	Category
McPherson and Burns, 2007 [48]	Gs Invaders: Assessing a computer gamelike test of processing speed	Space Code	Testing
McPherson and Burns, 2008 [89]	Assessing the validity of computer-gamelike tests of processing speed and working memory	Space Matrix / Space Code	Testing
Trapp et al., 2008 [90]	Cognitive remediation improves cognition and good cognitive performance increases time to relapse – results of a 5 year catamnestic study in schizophrenia patients	Xcog	Training
Gamberini et al., 2009 [91]	Eldergames project: An innovative mixed reality table-top solution to preserve cognitive functions in elderly people	Eldergames	Testing
Gamberini et al., 2010 [45]	Neuropsychological testing through a Nintendo Wii console	Wii Tests	Testing
Dovis et al.,2011 [92]	Can Motivation Normalize Working Memory and Task Persistence in Children with Attention-Deficit/Hyperactivity Disorder? The Effects of Money and Computer-Gaming	Megabot	Testing
Delisle and Braun, 2011 [93]	A Context for Normalizing Impulsiveness at Work for Adults with Attention Deficit/Hyperactivity Disorder (Combined Type)	Retirement Party	Testing
Prins et al., 2011 [50]	Does computerized working memory training with game elements enhance motivation and training efficacy in children with ADHD?	Supermecha	Training
Lim et al., 2012 [94]	A Brain-Computer Interface Based Attention Training Program for Treating Attention Deficit Hyperactivity Disorder	Cogoland	Training
Heller et al., 2013 [95]	A Machine Learning-Based Analysis of Game Data for Attention Deficit Hyperactivity Disorder Assessment	Groundskeeper	Testing
Hawkins et al., 2013 [47]	Gamelike features might not improve data	EM-Ants and Ghost Trap	Testing
Verhaegh et al., 2013 [96]	In-game assessment and training of nonverbal cognitive skills using TagTiles	Tap the Hedgehog	Both
Aalbers et al., 2013 [97]	Puzzling With Online Games (BAM-COG): Reliability, Validity, and Feasibility of an Online Self-Monitor for Cognitive Performance in Aging Adults	BAM-COG	Testing
Fagundo et al., 2013 [98]	Video game therapy for emotional regulation and impulsivity control in a series of treated cases with bulimia nervosa	Playmancer	Training
Anguera et al., 2013 [44]	Video game training enhances cognitive control in older adults	Neuroracer	Training
van der Oord et al., 2014 [99]	A Pilot Study of the Efficacy of a Computerized Executive Functioning Remediation Training With Game Elements for Children With ADHD in an Outpatient Setting	Braingame Brian	Training
Brown et al., 2014 [100]	Crowdsourcing for cognitive science--the utility of smartphones	The Great Brain Experiment	Testing

Tong and Chignell, 2014 [51]	Developing a Serious Game for Cognitive Assessment: Choosing Settings and Measuring Performance	Whack-a-mole	Testing
Katz et al., 2014 [101]	Differential effect of motivational features on training improvements in school-based cognitive training	WMTrainer	Training
Dunbar et al., 2013 [73]	Implicit and explicit training in the mitigation of cognitive bias through the use of a serious game	MACBETH	Training
Lee et al., 2013 [102]	A Brain-Computer Interface Based Cognitive Training System for Healthy Elderly : A Randomised Control Pilot Study for Usability and Preliminary Efficacy	Card-Pairing	Training
Miranda and Palmer, 2013 [49]	Intrinsic motivation and attentional capture from gamelike features in a visual search task	Visual Search	Testing
Atkins et al., 2014 [103]	Measuring Working Memory Is All Fun and Games A Four-Dimensional Spatial Game Predicts Cognitive Task Performance	Shapebuilder	Testing
Dörrenbächer et al., 2014 [52]	Dissociable effects of game elements on motivation and cognition in a task switching training in middle childhood	Watermons	Training
McNab and Dolan, 2014 [104]	Dissociating distractor-filtering at encoding and during maintenance.	The Great Brain Experiment	Testing
O'Toole and Dennis, 2014 [105]	Mental Health on the Go: Effects of a Gamified Attention-Bias Modification Mobile Application in Trait-Anxious Adults	ABMTApp	Training
Tenorio et al., 2014 [106]	TENI: A comprehensive battery for cognitive assessment based on games and technology	TENI	Testing
De Vries et al., 2015 [107]	Working memory and cognitive flexibility-training for children with an autism spectrum disorder: a randomized controlled trial	Braingame Brian	Training
Dovis et al., 2015 [108]	Improving Executive Functioning in Children with ADHD: Training Multiple Executive Functions within the Context of a Computer Game. A Randomized Double-Blind Placebo Controlled Trial	Braingame Brian	Training
Kim et al., 2015 [109]	Effects of a Serious Game Training on Cognitive Functions in Older Adults	Smart Harmony	Training
Manera et al., 2015 [110]	'Kitchen and cooking,' a serious game for mild cognitive impairment and Alzheimer's disease: a pilot study	Kitchen and Cooking	Both
Ninaus et al,. 2015 [32]	Game elements improve performance in a working memory training task	GAME	Training
Tarnanas et al., 2015 [111]	On the comparison of a novel serious game and electroencephalography biomarkers for early dementia screening	VAP-M	Testing

2.4 Results

2.4.1 Why have researchers used gamification?

I searched each article for reasons as to why researchers had chosen to use a gamified task. These reasons were then grouped into seven categories. Some authors listed multiple reasons for gamifying their approach, whereas others gave no motivations at all. Appendix A (pg141) provides details of which gamified tasks fell into each category.

To increase participant motivation:

Although I assume that increasing participant motivation was a goal for every study in this review, I found sixteen studies that explicitly used gamified tasks to measure or train cognition in a more motivating way, and the majority (10/16) of these studies were testing studies. Cognitive tests are typically a one-off measure, and re-playability is not a requirement. These gamified tasks made use of simple game-archetypes (such as space invaders or whack-a-mole), grafting them onto an existing cognitive task with the goal of encouraging self-motivation, improving participant enjoyment and even reducing test anxiety [48].

To increase usability/intuitiveness for the target age group:

Eleven tasks were gamified specifically to enhance appeal with a given age group (Appendix B - pg141). The authors of these studies hypothesised that a more intuitive interface could prevent boredom and anxiety in the target age group, which might damage motivation and concentration on the tasks at hand. Six gamified tasks were designed to be suitable for the elderly, who may not be familiar working with a mouse and keyboard [112]. Five other gamified tasks were aimed at young children, and re-framed the cognitive test as a game, in order to test the children under optimal mental conditions [106].

To reduce attrition:

Commercial gamification is often used to create long-term interest around a user experience, product or event [84]. Similarly, I found eleven studies (eight gamified tasks) that used gamification to reduce participant dropout rates over a protracted testing or training programme. Many of these gamified tasks were used in an unobserved, non-laboratory setting, and game design elements were needed to make the task more appealing and less burdensome. Task sessions were kept as short as possible to increase likelihood of completion [97,100]. For example, *The Great Brain Experiment* aimed to keep task sessions below five minutes in duration, and found that the shortest task, the stop-signal task (SST), was the most popular mini-game.

To investigate the effects of gamified tasks:

Many of the studies in this review investigated the effects of gamified tasks; however, only five studies explicitly stated that their aim was to assess the motivational and cognitive effects of gamification. Three of the six gamified tasks were very simplistic, with only a few specific game elements and carefully designed non-game controls, designed to make the effects of the game elements on the data as apparent as possible.

To stimulate the brain:

Six studies cited evidence that playing video games can be cognitively beneficial as a key factor in their decision to gamify. The past decade has seen considerable investigation into the cognitive effects of video gaming. Early findings were positive, with video gamers outperforming non-gamers on tests of working memory (WM) [113] visual attention [114,115] and processing speed [116,117]. Subsequent evidence has been less encouraging [68], but it is understandable that researchers hoping to train cognition are still keen to include gamified elements in their tasks.

To increase ecological validity:

Cognitive training has often suffered from a lack of transferability [46,58,118]. While participants may improve at the training task, these improvements do not generalise to the real world. In a similar vein, cognitive tests have been accused of being ecologically invalid [119]. A potential solution is to make tasks more realistic; and these tasks inevitably become gamelike as 3D graphics, sounds and narrative are added. I identified six studies that used gamified tasks to test cognition in an engaging close-to-life environment or enhance transferability of learned skills. As Dunbar and colleagues explain [73], games are uniquely suited to some forms of cognitive training as they give the player freedom to make choices and experience feedback on the effects of those choices, in other words they provide opportunities for experiential learning [120].

To increase suitability for the target disorder:

Gamified tasks may also be more appealing to patients with certain clinical conditions. Specifically, I found six studies (covering four gamified tasks) designed for people with Attention Deficit/Hyperactivity Disorder (ADHD). It is commonly reported that ADHD patients are compulsive computer game players [93]. Furthermore, patients with ADHD react differently from controls to rewards and feedback: they prefer strong reinforcement and immediate feedback, as well as clear goals and objectives, all of which can easily be delivered in a gamelike environment [50,92,121].

2.4.2 What cognitive domains has gamification been applied in?

I found comparable numbers of gamified *tasks* used for cognitive testing (17) and training (13), with one game that could be used for both testing and training (Appendix C - pg141). The numbers of *studies* investigating cognitive testing (17) and training (15) was also very similar.

Figure 2.2 shows the cognitive domains addressed by the tasks in this review. WM was the most commonly tested domain. This is likely due to its ease of testing and the fact that WM deficit is a common symptom in many disorders. General executive function, attention and inhibition were also commonly tested domains. Many gamified tasks were found to measure several cognitive domains and/or general EF, highlighting the difficulty of examining the effects of gamification on specific cognitive functions. With respect to training games, again WM was a popular target, closely followed by EF and inhibitory control training. I saw a smaller overlap in the domains covered by cognitive training tasks, they typically focussed on one or two domains exclusively while test batteries tended to assess a wide range of cognitions.

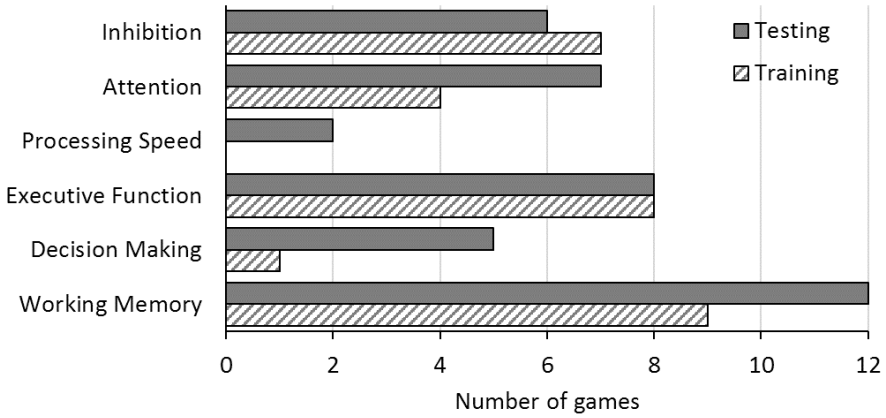


Figure 2.2 Cognitive domains addressed by gamified tasks in the review, shown separately by training and testing games.

2.4.3 What game design elements have been used?

Figure 2.3 shows a breakdown of the game design elements used by gamified tasks in this review. I coded game elements only if they were directly described in the paper or if a figure indicated their presence. Most tasks made use of 2D graphics, and some game elements, such as score, sound effects, theme and positive feedback appeared in almost every game. Figure 2.4 shows screenshots from games included in this review, displaying the wide range of game types and design elements used. For example, the screenshot in the top left in taken from *The Great Brain Experiment's* SST. A game which featured 2D graphics, background music, competition, minigames, non-interactive characters, points, sound effects and a strong theme.

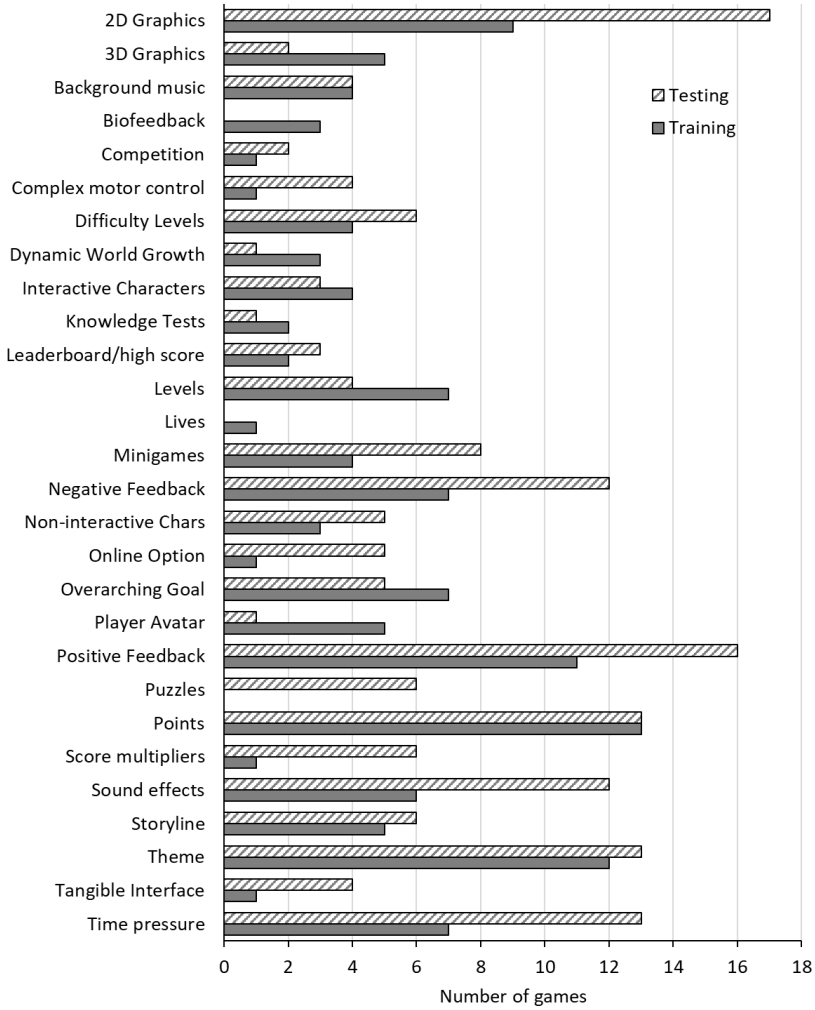


Figure 2.3 Bar chart showing the number of gamified tasks in the review that made use of each game design element. Game elements were only coded if they were described in the task's associated paper or if a figure clearly indicated its presence. Shown separately by testing and training.

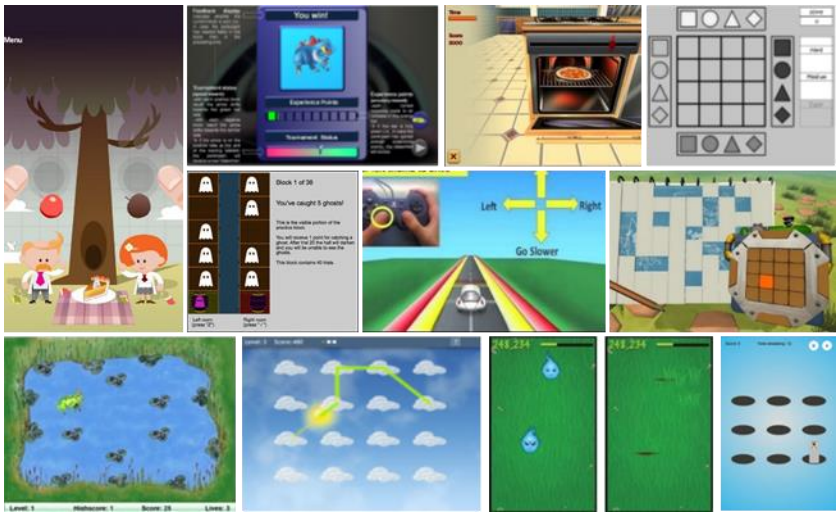


Figure 2.4 Selection of images of gamified tasks included in this review. From top to bottom, left to right: The Great Brain Experiment, Watermons, Kitchen and Cooking, Shapebuilder, Ghost Trap, Neuroracer, Braingame Brian, WMTrainer, BAM-COG, ABMTApp and Whack-a-mole.

2.4.4 What theory has guided gamification's application?

Of the 33 articles I reviewed, 25 made no reference to any motivational theories. The remaining eight papers touched on three theories: Flow Theory [122], Malone's theory of Intrinsically Motivating Instruction [123] and Self-Determination Theory [124]

Flow Theory

Conceived by Csikszentmihalyi in 1975 [122], flow is a temporary psychological state that may occur when performing an activity. While in a state of flow, we lose track of time, lose our sense of self, and become completely focused on the task at hand. This provides a sense of "inner clarity and deep satisfaction", causing the activity to become *autotelic*, that is, enjoyable for its own sake [122]. Flow is most likely to be induced by activities which have clear goals, deliver immediate feedback, provide the user with a sense of control, and most importantly, are challenging to a degree that perfectly matches the user's skill [125]. These conditions are commonplace in video games, and it has been argued that games are popular solely because they provide easy access to flow states [27,126–128]. I saw Flow theory applied in two ways: firstly, it was used as a framework for assessing whether gamification made the task autotelic [91]. This was done using questionnaires designed to quantify flow (such as the Dispositional Flow Scale and Flow State Scale) [129–131] and by using unobtrusive behavioural measures and physiological correlates [49,132]. Secondly, flow theory was used to guide the design of the gamified task (i.e., game elements were selected to provide goals, feedback and challenges with the hope of inducing a flow state) [49].

Malone's Theory of Intrinsically Motivating Instruction

Drawing on Csikszentmihalyi's work, Malone's theory of Intrinsically Motivating Instruction was first conceived in 1980 [76]. Malone rooted his theory in educational games, and used the term *intrinsically motiving* much as Csikszentmihalyi used the term *autotelic*: to describe a task that is innately enjoyable and appealing. Malone's theory sets out a framework for designing intrinsically motivating tasks, by evoking three things: challenge, fantasy and curiosity [123]. He argues that challenge is facilitated through the inclusion of goals, uncertain outcomes and rewarding feedback. That fantasy can be conjured through narrative framing, graphics and sound combined (i.e., a theme), and that fantasy provides a metaphor for the purpose of the task. By using a metaphor, the user can intuitively understand their goal and how to approach it. Finally, it is argued that curiosity is evoked through the provision of an optimal level of information complexity [132], in other words, the task should be simple enough that one can begin completing it, but it should hint at an underlying complexity: something waiting to be discovered if the user continues to engage [123]. In this review, Malone's theory was used only

by Katz and colleagues [101] as the basis for introducing fantasy to their WM training task through the use of themed graphics.

Self-Determination Theory

The most modern theory of motivation seen in this review is self-determination theory (SDT) [124]. SDT is concerned with the interplay between two driving forces for motivation: *intrinsic motivation* which relates to internal motivations such as interest, curiosity, or enjoyment; and *extrinsic motivation,* which describes motivation arising from external factors such as money, power or social pressure. SDT posits that humans have three basic psychological needs: autonomy, competency and relatedness, and that activities which help us to meet these needs are intrinsically motivating [133,134].

Autonomy relates to our ability to control the direction of the activity and our part in it: are there multiple ways to address problems? Is our taking part in the activity voluntarily? Competency refers to our need to perform well at an activity, to understand our level of performance and to be able to improve upon it. Finally, relatedness refers to our need to be connected to other people, both in kinship and competition. SDT was formulated on the basis of empirical studies and has stood the test of time. It has also seen uptake in the video-games literature as a tool for exploring their appeal [135,136]. In this review, SDT was referenced by five studies [49,52,101,110,137] and was primarily used to guide game development. For example, *Watermons* [52] provided clear feedback to players on their performance on the task, thus meeting the competency need. It also provided several different paths that the player could take as they completed the task, such as optional levels and a variety characters to play as: these choices were intended to facilitate player autonomy.

2.4.5 How have researchers measured engagement?

Seven studies measured participant engagement with the task. Six of these assessed participants' subjective experience of the task using self-report questionnaires [45,47,49,50,96,101]. Most of these questionnaires were created specifically for the study, with no standard questionnaire emerging. Items included "How much did you enjoy the task?", and "Would you be willing to recommend this task to a friend?", and participants responded using Likert and visual analogue scales (VAS). In some cases, these questionnaires were guided by motivational theory, for example Katz and colleagues [101] drew questions from the Intrinsic Motivation Inventory (IMI): a questionnaire commonly used in SDT research [46,138].

Two studies assessed engagement using objective measures of behaviour [50,52]. Prins and colleagues [50] asked participants to wait alone in the laboratory for 15 minutes after the task

was complete. During this period, they were allowed to: read, sit in silence, or continue playing the task. Researchers monitored player behaviour during this break to see whether either the gamified or non-gamified task was engaging enough that participants would *choose* to play it. In a similar way, Dörrenbächer and colleagues paused their training task five times to ask participants whether they'd be willing to complete an additional block of trials [52]. The non-game control presented this as an opportunity to "work through another block", while the gamified variant presented this as an opportunity to train the player's virtual pets. The task was designed such that regardless of how many 'optional' blocks were chosen, all participants trained for the same amount of time.

2.4.6 Does gamification work?

All study-authors were enthusiastic about their use of gamified tasks, although given the diversity of study aims; this does not mean that all gamified tasks worked as expected. Where reported, subjective and objective measures of participant engagement were positive. I identified 19/33 studies that compared a gamified task directly against a non-gamified counterpart, and these studies shed light on the specific effects of gamification on testing and training tasks (Table 2.2).

Table 2.2 Summary of evidence from studies which directly compared a gamified task to a non-game counterpart. Evidence was assessed only with respect to test validity or training outcomes, not participant engagement with the task.

Evidence for a successful test validation or an improved training effect?	**Studies**	**Count**
Yes	[32,45,47,48,50,96,97,100,103–105]	11
Mixed	[52,89,92,106,111]	5
No	[49,93,101]	3

Task Validity

Most gamified tests were validated successfully. Five games were of note: *Wii Tests* – four simple cognitive tests aimed at older people and delivered through the Nintendo Wii [45]. *Shapebuilder* – a colour and shape categorising game designed to measure WM and EF [103]. *The Great Brain Experiment* – a suite of graphically rich minigames testing a variety of domains and comparing participants 'score' against other users [100,104]. *BAM-COG* – four themed puzzle games developed to measure WM, visuospatial short-term memory, episodic recognition memory and EF [97]. *Tap the Hedgehog* – a tactile boardgame based on the corsi-block tapping task [96]. These games produced output measures that correlated fairly well with their non-gamified counterparts (full details of included gamified tasks can be found at goo.gl/PYVrH7). Validation studies varied in their design, and some studies reported complex

correlations between gamified and non-gamified tasks with multiple outputs. Sample correlations from some of the simpler validation studies suggest inter-task correlations of .45 to .60 [97,103,104,139]. Broadly speaking these were well-designed and well powered-studies, and together they provide encouraging evidence that cognitive tests can be gamified and still be useful as a research tool.

Some studies found that their gamified tests were correlated with measures of multiple cognitive domains; in other words, they were mixed-domain measures. Mixed-domain measures have upsides and downsides: they are good at detecting the presence of cognitive deficits but are less suitable for identifying specific deficits. Accordingly, they are less useful for developing models of cognition or pinpointing deficits to specific brain regions. Use of exploratory factor analysis showed that *Whack-a-Mole's* primary output measure was correlated with two of the three executive functions of interest: inhibition (r=.60, p<.001) and updating (r=.35, p<.05) [51]. *Space Code* [48,89] had similar problems. The initial study was successful, with *Space Code's* output measure correlating well with a conventional measure of processing speed (r=.55, p<.001). However, a second article detailing two experiments which aimed to replicate the previous finding found that *Space Code's* correlations with measures of WM, visuospatial ability and processing speed were not stable [89]. The fact that *Space Code* was thought to be a pure measure in one study and then was found to be mixed-domain in the next highlights that designing gamified cognitive tasks is difficult, and multiple, well-powered validation studies may be required to ensure a task is measuring what is intended.

Gamification also has the potential to invalidate a task. For example, *Retirement Party* was compared against the Continuous Performance Task-II in healthy controls and adults with ADHD. The Continuous Performance Task-II detected more commission errors from the ADHD adults as expected (M=56, SD=13 vs M=46, SD=10), but *Retirement Party* did not (M=14.4, SD=5.8 vs M=13.2, SD=4.3): this likely invalidates the game as a diagnostic tool for ADHD. However, Delisle and Braun [93] discuss the possibility that *Retirement Party* may have detected no deficit in ADHD patients as the nature of the task was such that there was no deficit: the highly structured and feedback-rich multitask environment may have normalised the ADHD patients' usual inattention. Such a performance boost resulting from game design elements is a disadvantage when performing a cognitive test but is likely desirable in a cognitive training scenario.

Studies directly investigating gamification

Several studies in this review focused specifically on adding game elements to cognitive tasks to investigate the resultant changes in data, enjoyment and motivation. Dovis and colleagues [92] studied whether different types of incentive could normalise ADHD children's performance on WM training. They found that regardless of incentive, ADHD children did not perform as well as healthy controls. ADHD children also experienced a decrease in performance over time, but this was prevented by both high financial incentive and gamification (*Megabot*). These results indicate that performance problems in ADHD training might be somewhat alleviated through the use of gamified tasks. This is further supported by Prins and colleagues study (*Supermecha* [50]) which found that ADHD children completed more training trials (M=199, SD=47 vs M=134, SD=34), with higher accuracy (69% vs 51%), when trained using a gamified WM training task as opposed to a non-gamified one. Children in the gamified condition were also opted to train for longer and enjoyed the training more. In a similar vein, *The Great Brain Experiment* [32] and *GAME* [104] both showed gamified tasks to be appropriate for measuring and training WM. With Ninaus and colleagues presenting evidence that gamification can improve overall participant performance in a WM training task [32], and McNab and Dolan showing that data collected from two very different gamified and non-gamified tasks could fit similar models of WM capacity [104].

In contrast, *WMTrainer* was assessed across seven different task variants which made use of different combinations of game elements [101]. They compared training improvement across conditions and found the greatest training effect in variants with minimal game design elements. The fully gamified variant had one of the shallowest improvement slopes and none of the task variants made any difference to subjective motivation scores. However, even the minimally gamified version still had simple graphics and displayed a player score at the end of the block. It is possible that even this minimal gamification was enough to increase motivation, and that adding 'distracting' game elements such as persistent score display may have a negative impact on performance by inducing stress or new cognitive demands [101].

One of the most theoretically driven tasks I reviewed, *Watermons* [52], found that gamification increased the effect of training (reducing RTs and task-switching costs) compared to a non-gamified version of the training. Participants were also more willing to train in the gamified condition; opting to complete more blocks.

Miranda and Palmer [49] used a visual search task with two forms of reward for fast and accurate responses: sound effects and points. They found that sound effects led to increased

RTs, potentially due to attentional capture, and did not improve ratings of subjective experience. Points appeared to have no effect on data and boosted subjective experience scores. These results highlight the delicate nature of designing gamified cognitive tests, since something as innocuous as a few sound effects had deleterious effects on participant performance.

Finally, Hawkins and colleagues [47] compared gamified variants of two decision making tasks against non-game counterparts. No difference between the data collected by the gamified variants and the non-game variants was found. Subjective ratings indicated that both variants of both tasks were equally boring and repetitive, but that the gamified variants were more interesting and enjoyable.

2.5 Discussion

This review identified seven reasons why researchers use gamification in cognitive research. These include the typical applications of gamification such as increasing long and short-term engagement with a task, but also more clinically related reasons such as making tasks more interactive in order to enhance the effect of cognitive training. Several studies aimed to reduce test anxiety and optimise performance in groups that traditionally dislike being tested, particularly electronically, such as elderly people and children. By hiding the test behind a novel interface and gameplay, the target audience might feel more comfortable.

I reviewed several gamified tasks aimed at training and testing people with ADHD, and overall these tasks appear highly engaging to users, in some cases even increasing the time spent training. Gamified tasks may be valuable for assessing ADHD patients as computer games are particularly appealing to them: with rapid rewards, immediate feedback and time-pressure being exactly the type of stimulus the ADHD brain craves [140,141]. The dopaminergic system is thought to be abnormal in ADHD [142,143]. However, it is thought that playing video games can facilitate the release of extra-striatal dopamine, which plays a role in focusing attention and heightening arousal [144,145]**;** this may improve player performance and motivation. Nevertheless, as Delisle and Braun discuss [93], we must be cautious that liberal use of game design elements does not reverse the very deficit we are hoping to measure.

One of the primary reasons that psychologists are keen to utilise gamification is to reduce the impact of low motivation and accordingly increase performance in research populations. Gamification might result in faster response times, increased accuracy, less RT variance etc, by reducing confounding caused by low motivation [47]. The results of [47,48,52,89], show that gamified tasks can be used to improve motivation while still maintaining a scientifically valid

task. However, Katz and colleagues [101] and Miranda and Palmer [49] highlight how difficult a balancing act this can be, with several game elements having unforeseen deleterious effects on performance. If gamified psychological tasks are to become common in the future, further research is required to disentangle the impact of specific game elements on task performance, as these studies have already begun to do.

2.5.1 Defining engagement

Studying the use of language in the reviewed articles revealed a variety of ways in which authors used the term *engagement*. Notably, the word rarely appeared on its own, instead being used as part of a word pair such as "engagement and enjoyment" [47,100] or "engagement and compliance" [105]. This cautious use of the term may be the result of its ambiguous meaning [146,147].

Engagement is a word found in many different contexts: from market research, to game design, to healthcare provision. It has many similar, yet subtly different, definitions, making its usage in scientific literature imprecise. On one hand, when we talk about users 'engaging' with a website: we are typically referring to *behaviour* [148]. For example, how often does a user return to the site? How long do they spend on the site? What content are they exploring? On the other hand, when we see an 'engaging' film, we refer to our *subjective experience* [149]. We enjoyed ourselves. We lost track of time. We became immersed in the film's story, etc. Of these two distinct meanings, the former has established itself in the language of digital behaviour change and much of the commercial sphere, while the latter definition is most at home in the fields of computer science and human computer interaction.

A recent systematic review by Perski and colleagues [150] formally integrated these two concepts into the following definition: Engagement is the extent (e.g., amount, frequency, duration, depth) of usage and the subjective experience (characterised by attention, interest and enjoyment). They conceptualise engagement as a multidimensional construct, with the user's subjective experience of the task driving the behavioural aspects of engagement.

Henceforth I will refer to the objective and behavioural aspects of engagement as **amount of engagement**, and the subjective, experiential and self-report aspects of engagement as **quality of engagement**.

2.5.2 The theoretical basis of gamification

Three theories of motivation were referenced in this review: Flow Theory [122], Malone's theory of Intrinsically Motivating Instruction [123] and SDT [124]. I argue that despite conceptual differences between Malone's *intrinsically motivating activity* and SDT's

intrinsically motivating activity, and flow theory's *autotelic* activity, all these theories posit a causal link between the constituent elements of an activity and whether it is enjoyable for its own sake. Combining this hypothesis with the above definition of engagement [150] affords a logical third step: that the constituent elements of a task drive the users' quality of engagement (subjective experience), and that quality of engagement in turn drives the users' amount of engagement. It is this hypothesis that forms the theoretical basis of gamification [38].

It is therefore curious that the majority of studies I reviewed made no mention of motivational theory whatsoever. Though the relationship between quality and amount of engagement is not yet fully understood [150], there is evidence from both flow theory and SDT on the relationship between task components and subjective experience [127,151]. I assume that many of the authors in this review implicitly assumed the above hypothesis, since the very act of gamifying a task implies a causal chain from game elements through to increased amount of engagement.

2.5.3 Differences between training and testing tasks

Gamified training tasks typically contained many game design elements and were similar in appearance to commercial video games (Figure 2.5). Cognitive training normally requires several sessions to be effective, and as a result, training tasks need to be engaging enough to play for many hours. 3D graphics were quite prevalent, as was the use of avatars, points, levels and dynamically growing game worlds (Figure 2.3). Long-term goals which had to be completed over repeated sessions were also common and served to sustain amount of engagement over a long period of time.

In contrast, gamified tests were simpler, predominantly using 2D graphics, sound effects, score and theme to create the appearance of a game. Several gamified tests simply presented themselves as "puzzles" which the participant had to complete. Tasks of this nature represent gamification at its simplest, but they were well received by users, implying that minimal gamification is better than no gamification.

In general, gamified cognitive tests typically used fewer game design elements (Figure 2.5) than gamified training tasks and restricted themselves to mostly superficial changes such as the introduction of points, themed graphics, and sound effects. This is likely due to the constant tension between creating an engaging task and the risk of undermining the task's scientific validity: including unknown game elements might have deleterious effects on the data collected.

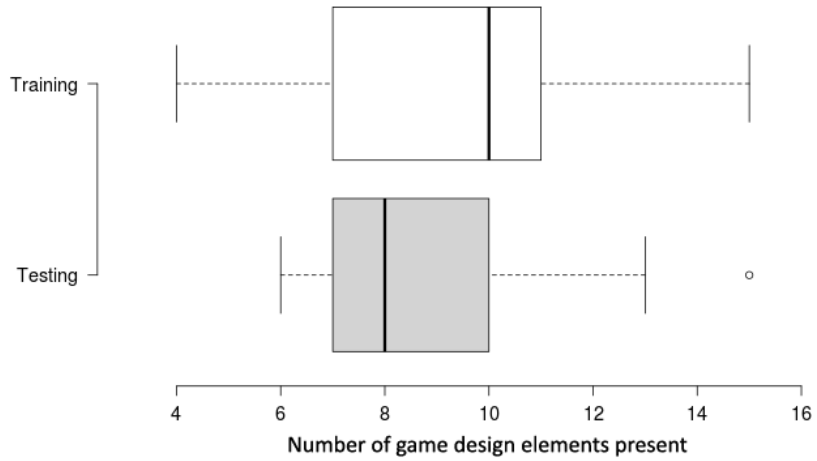


Figure 2.5 Total number of game design elements present in a task, shown separately training and testing tasks.

2.5.4 Validating gamified tasks

I found heterogeneous standards for validating gamified tasks. Typically, gamified cognitive tests were validated rigorously, using correlation with similar cognitive tasks and factor analysis to determine whether they were performing as expected. Many training studies investigated a gamified task only, meaning the effect of gamification cannot be dissociated from the effect of the intervention. Sample sizes were quite small (Figure 2.6) in many studies, and there was little consideration of statistical power when sample sizes were decided upon, with only 5/33 articles describing a power calculation. Gamified cognitive tests are novel scientific instruments and must be validated as such. Small pilot studies, followed by larger validation studies including assessment of test-retest reliability, and internal and external validity of the measures taken by the game are needed [152]**.** Regarding cognitive training, ideally gamified training should be treated as an intervention and so the current gold standard of a blinded randomised control trial is appropriate [153].

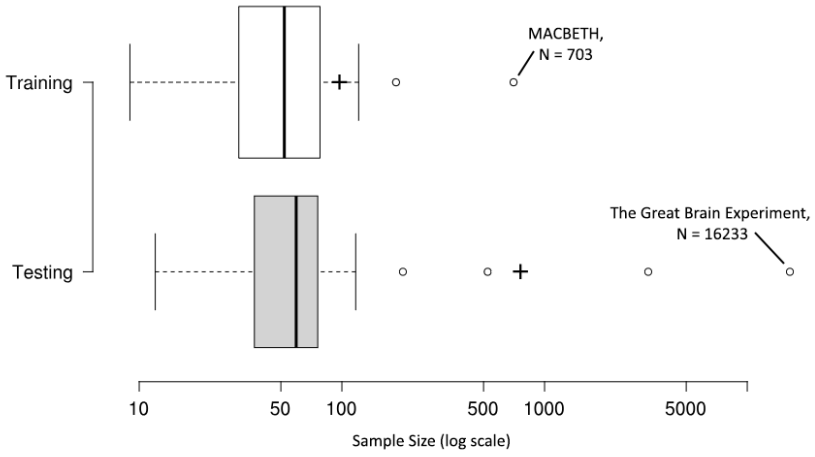


Figure 2.6 Boxplots of study sample sizes, shown separately by training and testing studies.

2.5.5 Limitations

One limitation of this review is the possible subjectivity in the selection process. Gamification in psychology is a rapidly growing field, hence I decided to focus specifically on 'gamified cognitive training and testing'. This resulted in some articles being excluded on the subjective basis of 'not being cognitive' or 'not being gamified'. Nevertheless, to counteract this subjectivity, articles were only included in the review on consensus from both reviewers (JL and EE), and a 20% selection check was performed by EE on articles from the title-screening stage onwards. An additional consideration is that many of the studies reviewed were of a preliminary nature, and as such the findings reported here should be considered tentative.

2.6 Chapter Summary

In this chapter, I systematically reviewed the literature on gamified cognitive tasks. I used a broad search strategy, selecting primary research articles which had validated a novel gamified task to measure or train cognition, and were published between January 2007 and October 2015.

Researchers used gamification for a variety of purposes, including increasing both quality and amount of engagement. Gamification was used to reduce test anxiety, increase training transfer and increase task suitability for specific age groups and patient groups. A wide variety of gamified tasks were reviewed, from simple 2D puzzles through to fully interactive 3D training games (Figure 2.4)

I found equal numbers of training and testing tasks, and the mostly commonly addressed cognitive domain was WM, followed closely by general EF, attention and inhibition. It has been suggested that gamifying a cognitive task might result in higher quality data or more effective training, either by reducing between-subject noise, attenuating motivational differences or by improving participant performance [47], but my review found no evidence to support this.

In two-thirds of studies gamified tasks were directly compared against non-gamified tasks; and evidence was generally in favour of gamification (Table 2.2). Gamified tests were found to be valid, and gamified training improved training outcomes more than control. However, there were cases where gamification worsened participant performance [49,89,101] or had mixed effects [52,89,92,106,111]. Furthermore, one third of studies lacked appropriate controls to distinguish the effect of gamification from the test or intervention. Study designs were heterogenous, and sample sizes were typically small (Figure 2.6). Many studies introduced multiple game design elements into a task at once, making it difficult to determine which game elements were driving which effects (Figure 2.5).

I surveyed the way in which researchers use the term *engagement* and sourced a more rigorous definition. Henceforth I will refer to the subjective, experiential, and self-report aspects of engagement as **quality of engagement**, and the objective and behavioural aspects of engagement as **amount of engagement.** I also investigated methods used to assess participant engagement and found that self-report measures of quality of engagement were far more common than measures of amount of engagement. Most of studies I reviewed stated that gamification was well received, and measures of engagement were universally positive. Despite concerns that some commonly used game elements might reduce participant motivation [154] (such as by having a visibly low score [155]), I found no evidence that this was the case. However, the recurrent investigation of only one dimension of engagement reduces the certainty of these findings.

I assessed the motivational theories, and the relative lack thereof, that underpinned the reviewed studies: briefly overviewing Flow Theory [122], Malone's theory of Intrinsically Motivating Instruction [123] and SDT [124]. These theories all posit a causal link which, when combined with the above definition of engagement, results in the hypothesis that underpins the concept of gamification. That the constituent elements of a task drive the users' quality of engagement, and that quality of engagement in turn drives the users' amount of engagement.

In conclusion, this review raises several interesting directions for research. Promisingly, the evidence suggests that it is *possible* to develop valid gamified tests and effective gamified training, though this was not the case for all articles reviewed. I also saw evidence that gamification was effective for increasing engagement, though mostly with respect to quality of engagement.

Nevertheless, the field of gamified cognitive tasks is still in its infancy, and more rigorous study designs with larger sample sizes are needed. Furthermore, the studies I reviewed provide little insight into the relationship between individual game design elements, engagement, and cognitive processes. The next chapter documents an empirical study I conducted to explore this relationship further.

Chapter 3: The effects of points and theme on test data and quality of engagement (Experiment 1)

This chapter is based on my publication in PeerJ [156].

3.1 Chapter Aims

This chapter documents my first empirical study into the effects of gamification of cognitive data and participant quality of engagement. I had four aims, namely to:

1. Develop an experimental methodology to answer the research questions of this thesis
2. Investigate the effects of individual game elements on the primary cognitive outcome measures of the Go/No-Go task (GNG)
3. Investigate the effects of individual game elements on quality of engagement
4. Investigate the impact of testing online compared to testing in the laboratory

3.2 Introduction

In the previous chapter I reviewed the literature on gamified cognitive tasks and identified a shortage of research examining the effects of *individual* game elements. In many cases, gamified tests contained multiple game elements, making it difficult to determine causal links between game elements and effects on cognitive data or quality of engagement. My review found mixed, though generally positive, evidence for the effects of gamification on cognitive test data. On one hand, gamification could be a benefit [32,92]. Low motivation to complete experimental studies may have negative effects on data quality, adding noise and leading to suboptimal performance [21], and gamification may be able eliminate this problem. On the other hand, there was also evidence that gamification can worsen participant performance [49], potentially by introducing new task demands. The evidence suggested that gamified tasks typically improved participant quality of engagement [48,51,52].

As described in Section 1.1, web-based testing has rapidly risen in popularity over the last decade; enabled, in no small part, by Amazon Mechanical Turk (www.mturk.com) [6]. MTurk is a 'work marketplace' which allows users to sign up, complete small online tasks and receive reimbursement for their time. While MTurk is mostly used for non-research purposes, it has grown popular in the behavioural sciences because it enables the testing of large numbers of people in a very short time. Studies investigating the comparability of data from laboratory and online versions of tasks have reported mixed findings [1,3,157,158]. These differences may arise from a number of factors, including: differences in the population sampled (with online participants tending to be older than those recruited through traditional methods), differences in hardware used, the suitability of the remote environment for concentration and reduced

motivation due to lack of experimenter presence [4]. Engaging participants in online studies is particularly challenging [8,9], making web-based testing an obvious application area for gamification.

In this study, I had two reasons for comparing data collected online and in the laboratory. Firstly, I wanted to investigate differences in cognitive data and check whether web-based testing would be a suitable methodology for the remainder of my research. Secondly, I wanted to see if gamification had the same effect on engagement across both sites.

The goal of this study was to compare three variants of a GNG, delivered both in the laboratory and online using Xperiment, a web-based platform for psychological experiments (www.xperiment.mobi). Online participants were recruited and paid using MTurk, but both online and laboratory participants used Xperiment's specialist software to complete the experiment itself. The three task variants included: one variant where participants were rewarded with points for performing optimally (points variant), one where the task was themed as a cowboy shootout (theme variant), and a standard GNG as a control condition. These task variants were chosen because my review showed points and theme to be among the most common game design elements used in gamified cognitive tasks (Figure 2.3).

Points and scoring systems have a long history in all types of games. Though seemingly simple, their mechanisms of motivation are not fully understood, and evidence suggests they serve multiple functions. A full review of the psychological effects of points is beyond the scope of this thesis, but I will touch on two keys aspects here.

Firstly, points may serve as a reward in and of themselves. A common assumption among games designers is that dopamine is *the* brain chemical responsible for pleasure and enjoyment, and that points (and other in-game rewards) provide players the 'dopamine hit' they are seeking [159]. Though this interpretation is an oversimplification, there is evidence from the animal behaviour literature of an association between reward (sucrose water, food pellets, amphetamine, etc) and dopamine release [160–163]. Similarly, In humans, positron emission tomography (PET) scans have shown increased striatal-dopamine release as a result of playing a reward rich video-game [164]. That said, dopamine's relationship with reward and has not been fully untangled [165]; and there's evidence of associations between dopamine and hedonic reward [166], motivation for reward [167], learning how to achieve reward, and more (see [168] for a review).

Secondly, points may serve to quantify feedback on player performance. Depending on the implementation, feedback valance (positive or negative) might be indicated by the magnitude of the points score (high or low scoring trials), or by the sign of the points score (gaining or losing points). In both cases, points may help the player to improve their performance by indicating which strategies are effective (positive feedback) and which strategies are not (negative feedback) [169]. Theoretically speaking, both Self-determination Theory (SDT) and flow theory, posit that performance-related feedback helps to meet the competency need and induce a flow state [122,134]. But empirical evidence suggests that the effects of feedback on motivation, performance and effort are not so simple (see [169,170]).

3.2.1 Hypotheses

When designing the gamified task variants, I aimed for similarity to the non-game control in the hope of minimising any impact on cognitive data. I therefore hypothesised no difference between median RT or mean No-Go accuracy between any of the task variants but had no expectations regarding differences in Go accuracy between the task variants. Based on the findings of Chapter 2, I hypothesised that participants would have a higher quality of engagement with the theme and points variants compared to the non-game variant. I also hypothesised that participant quality of engagement would be lower in the online condition.

3.3 Methods

3.3.1 Design and Overview

I used a 2 × 3 between subjects design, with task variant (non-game, points, theme) and test location (laboratory, online) as factors. The dependant variables of interest were RTs on Go trials, Go trial accuracy, No-Go trial accuracy and quality of engagement. The study was pre-registered on the Open Science Framework (osf.io/va547).

3.3.2 Participants and Procedure

Participants were recruited through existing email lists and poster advertisements around the University of Bristol. These participants were tested in the laboratory and received either course credit or £3 in compensation for their time. A second group of participants were recruited through MTurk; they received payment of $1.50. All participants required to be older than 18 years of age, not have a diagnosis of Attention Deficit/Hyperactivity Disorder and not be colour blind. Once enrolled, participants were randomly assigned to one of the three task variants. Since testing site (laboratory or online) was determined by the participant's method of signup, the groups were not matched.

Study sessions lasted approximately 15 minutes. Each participant took part in only one task variant. Participants confirmed they met the inclusion criteria and provided consent using an online form. Demographic information was collected on the participant's age, sex, ethnicity, level of education and the number of hours they spent playing video games per week. Participants then began the study, instructions for the task were displayed, and the GNG task was delivered, followed by the questionnaire and finally a debrief screen. Participants were free to withdraw from the study at any point by simply closing the browser window, this would result in no data being saved. Ethics approval was obtained from the Faculty of Science Research Ethics Committee at the University of Bristol (ref: 22421). The study was conducted according to the revised Declaration of Helsinki, 2013 [171].

3.3.3 Materials

Online and Laboratory platforms

To eliminate task differences caused by variations in delivery platform, I used Xperiment to host both the lab and the online version of the task. Xperiment is an online experimental platform which has been shown to collect comparable data to other, offline, test software [172,173]. Laboratory participants were seated in a computer cubical while they completed the task and the questionnaire via the Internet. They used a PC with a mouse and keyboard to complete the task. MTurk participants accessed the same experimental software, but via their own PC or laptop.

Go/No-Go Task

Response inhibition, i.e., the ability to stop or withhold a motor response, is a key feature of executive control [174]. The GNG is one of two tasks commonly used to assess response inhibition (the other is the stop-signal task, see Section 5.3.3): it comprises a reaction-time task with a fixed set of no-action stimuli. It tests inhibitory control by repeatedly presenting stimuli to which the participant must respond rapidly, while occasionally presenting stimuli to which the participant must avoid responding.

I decided to use a GNG task on the basis that it was a natural fit for gamification to a cowboy shootout theme. Several commercial arcade games (such as Duck Shoot) involve the player rapidly shooting specific targets (ducks) while avoiding others (baby ducks). Such games are GNG tasks in disguise, and provide a good metaphor for the actions required by the GNG task [123]. The GNG task also has additional qualities that make it suitable in this context. Firstly, it is a commonly known and commonly used cognitive task. Secondly, it records RT, allowing me to detect potential slowing effects caused by gamification. Thirdly, it is a relatively simple task,

with simple outcome measures, and therefore served as a first step on the road to investigating gamified cognitive tasks in more detail.

I developed my own GNG task for use on the Xperiment platform, based on the tasks used by Benikos and Bowley and colleagues [175,176], but with custom game elements for each variant. Each trial began with a fixation cross displayed in the middle of the screen, 500 ms later a picture appeared in the centre of the screen and remained for 600 ms. On Go trials the participant had to respond to the stimuli as fast as they could by pressing the spacebar within this 600 ms window. In No-Go trials (signalled by the image content) they simply had to withhold their response. Each trial was followed by a variable inter-trial-interval of 500-1000 ms. If the participant responded incorrectly, the inter-trial interval was replaced by a feedback screen, failed No-Go trials resulted in a red cross being overlaid on the stimuli, while incorrect no-responses were followed by "Too slow" written in red text.

The task consisted of 5 blocks of 60 trials each. Between each block a pause screen was displayed and the participant had to wait for 10 seconds. Each block contained 5 sub-blocks of 12 trials, and each sub-block consisted of 9 Go trials and 3 No-Go trials, in a randomised order. In total, the task contained 75 No-Go trials (25%) and 225 Go trials (75%) and took around 11 minutes to complete. GNG tasks vary widely in their design, but using 25% No-Go trials is similar to several other studies [175,177–179].

Non-game Variant

The non-game variant used a stimulus set consisting of a diverse range of 20 everyday objects: 15 green and 5 red. Go trials used the green object, and No-Go trials used the red objects (Figure 3.1A). I selected green and red objects to ensure that the non-game variant was as intuitive as the themed variant, as these colours are commonly associated with going and stopping [180]. Appendix D (pg142) shows the instructions presented to the participants.

Points Variant

The points variant was identical in structure to the non-game variant, except that a scoring system was added, based on that used by Miranda and Palmer [49]. The participant's score was displayed in the middle of the screen, to the left of the stimuli (Figure 3.1C). On each successful Go trial the participant earned points equal to $Bonus \times 0.1 \times (600 - \mathrm{RT})$. This bonus was a multiplier (2×, 4×, 8×….) which doubled every 5 trials but was reset to ×1 when the participant made a No-Go error. On a successful inhibition the bonus was not lost, but no points were awarded. This scoring scheme also fits with the findings of Guitart-Masip and colleagues [181], who found that subjects were much more successful in learning active

choices when rewarded for them, and passive choices when punished. The points awarded in the previous trial were displayed in the centre of the screen during the inter-trial interval. The instructions framed the task as a game (Appendix E - pg142).

Theme Variant:

The theme variant also used the same format as non-game variant, except with the addition of a theme designed to provide a narrative framework for the action required by the task (Appendix F - pg142). The participant was introduced to the task as a shooting game, where they were the sheriff of a small town and a group of criminals had holed up in a saloon and taken hostages. The GNG task proceeded as above but the stimuli were replaced with cartoon characters, with cowboys as Go targets and innocent civilians as No-Go targets (Figure 3.1E). Throughout each block a cartoon saloon graphic remained on the screen, with stimuli appearing in the doorway. When the participant pressed the response key a blood splat was overlaid onto the current stimuli for the remainder of the trial time. Feedback was presented in the inter-trial interval, as in the non-game variant. The stimulus set consisted of 15 cowboys and 5 innocent civilians.

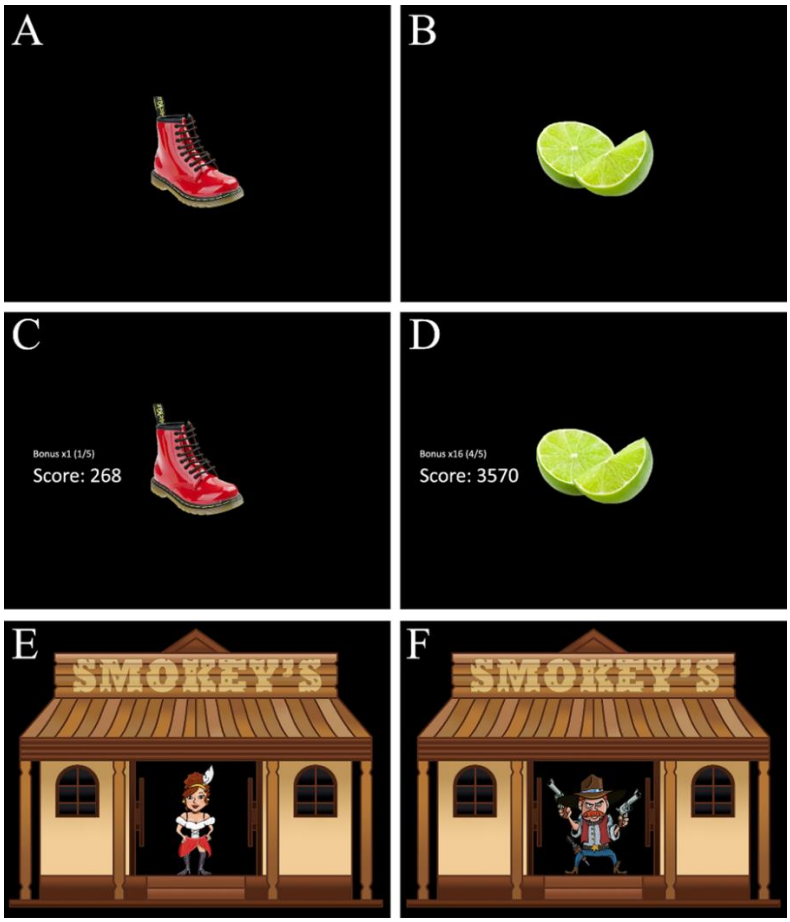


Figure 3.1 No-Go trial from the non-game variant. B: Go trial from the non-game variant. C: No-Go trial from the points variant. D: Go trial from the points variant. E: No-Go trial from the theme variant. F: Go trial from the theme variant.

Assessment of Quality of Engagement

After completing the task participants were presented with a brief questionnaire to assess their experience of the task. Following the approaches of Hawkins and colleagues., and Miranda and Palmer [47,49], 11 questions were selected :

1. How enjoyable did you find the task?
2. How frustrating did you find the task?
3. Was it difficult to concentrate for the duration of the task?
4. How well do you think you performed on this task?
5. How mentally stimulating did you find this task to be?
6. How boring did you find the task?
7. How much effort did you put in throughout the task?
8. How repetitive was the task?
9. How willing would you be to take part in the study again?
10. How willing would you be to recommend the study to a friend?
11. How intuitive did you find the pictures chosen for stop and for Go?

Participants responded using a continuous VAS, presented as a horizontal line with an appropriate label ("Not at all", "Very much") at each end and no subdivisions. Participants marked a point between these two labels using their mouse, which was recorded as percentage distance along the line. The questionnaire was delivered using the same Xperiment platform that delivered the tasks.

3.3.4 Dependent Variable Calculation

Go Trial Reaction Times

Go Trial RTs were summarised using the participant's median RT from all Go trials

Go Trial Accuracy

Go Trial Accuracy was summarised using the participant's mean accuracy (%) on all Go trials.

No-Go Trial Accuracy

No-Go Trial Accuracy was summarised using the participant's mean accuracy (%) on all No-Go trials.

Quality of Engagement

Quality of Engagement was measured by calculating the participant's mean VAS score from items 1-10 on the assessment of quality of engagement. Items 2,3,6 and 8 were reverse scored.

3.3.5 Bayesian Statistics

In the analyses below, where appropriate, differences between groups were assessed using *post-hoc t*-tests. Frequentist statistics are not ideal for testing equivalence [182,183], and so when there was no evidence of a difference between group-means, I used Bayesian *t*-tests and Bayes Factors (BF) to assess the evidence for no difference [184,185].

A Bayesian *t*-test produces a BF, which compares the evidence for two hypotheses. If the evidence favours one hypothesis over the other then the BF will reflect that, but if the evidence is equal for both hypotheses then the BF will imply that the data are insensitive [184,186,187] (Table 3.1). I used the Bayesian *t*-test procedure in JASP (jasp-stats.org), with a Cauchy prior width of 0.707. Setting the Cauchy prior width to 0.707 means that one hypothesis is "the effect size is zero" and the other is "the effect size is between -0.707 and 0.707". Although both hypotheses are centred on an effect size of 0, the former makes a stronger claim than the latter. As such, effect sizes which are not close to 0 are better represented by the latter hypothesis. A prior width of 0.707 was selected because it represents the expectation of a medium-large effect, thus weighting the BF against small effects and reducing the likelihood of a false positive.

Table 3.1 Interpreting Bayes factors (adapted from [186])

Hypothesis 0: The effect size is 0	**Strength of Evidence**	**Hypothesis 1: The absolute effect size is between 0 and X**
.33 ≤ BF ≤ 1	No support either way	1 ≤ BF ≤ 3
.1 ≤ BF ≤ .33	Positive	3 ≤ BF ≤ 10
.01 ≤ BF ≤ .1	Strong	10 ≤ BF ≤ 100
BF<.01	Decisive	BF>100

3.3.6 Statistical Analysis

Go Trial Reaction Times

To assess the effect of gamification on Go RTs I used a two-way ANOVA of Go Trial RT with task variant (non-game, points theme) and test location (laboratory, online) as between-subjects factors. I used box and whisker plots to compare Go RTs across the task variants.

Go Trial Accuracy

Accuracy data were handled similarly. I assessed the effect of gamification on Go accuracy using a two-way ANOVA of Go accuracy with between-subjects factors of task variant (non-game, points theme) and test location (laboratory, online). Again, I used box and whisker plots to compare Go accuracy across the task variants.

No-Go Trial Accuracy

I assessed the effect of gamification on No-Go accuracy using a two-way ANOVA of No-Go accuracy with between-subjects factors of task variant (non-game, points theme) and test location (laboratory, online). Again, I used box and whisker plots to compare accuracy across the task variants.

Quality of Engagement

I assessed differences in participants' quality of engagement scores both visually (as mean scores and individual items) and using a two-way ANOVA with task variant (non-game, points theme) and test location (laboratory, online) as factors.

3.3.7 Sample size determination

At the time of study design, no previous study had investigated differences in data produced by gamified and non-gamified GNG tasks, and I therefore had no previous effect size on which to base a sample size determination. My primary hypotheses estimated no effect of gamification on cognitive data, and so I based my sample size calculation of a positive effect of gamification on quality of engagement. To detect a medium effect (η^2=.05) of gamification on quality of engagement, I required a sample size of 297. I set this to 300 so I could divide the sample equally across the three task variants.

3.4 Results

The data that form the basis of these results are available from the University of Bristol Research Data Repository (10.5523/bris.1hjvqlpbtrk961ua9ml40bauie)

3.4.1 Characteristics of Participants

A total of 304 participants took part in this study, however four participants from the online group were excluded from the subsequent analyses because I did not record any responses from them for the duration of the GNG task. A further thirteen participants were excluded due to extremely poor Go accuracy rates (more than 4 inter-quartiles ranges away from the median).

Excluding these outliers, 287 participants took part: 84 in the laboratory (mean age = 21, SD = 4, 26% male) and 203 online (mean age = 35, SD = 11, 50% male). A chi-square test indicated that the number of male participants in the laboratory site was statistically different to the online ($X^2_{1,\ N=287}$ =14.012, p<.001). A t-test provided evidence for difference in ages between the laboratory group and online (t_{285}=16.35, p<.001), with the online participants being older on average. Participants who took part online reported slightly more hours spent playing computer games per week (median = "1 to 5") than those that took part in the lab (median =

"0")—there was evidence that the distributions of responses for both groups differed, with the laboratory group being skewed towards 0 (Mann-Whitney U = 3994, Online=203, Lab=84, p<.001). Online participants also reported higher levels of education (median = "Bachelor's degree") than those in the laboratory (median = "High School Graduate"), and there was evidence that these distributions differed, with 83% of the laboratory group being high school graduates and the online group being a relatively even split between high school graduates and university graduates (Mann-Whitney U = 5330, Online=203, Lab=84, p<.001). However, given that most laboratory participants were undergraduates, they would be equally educated within a few years. Ethnicity also differed between sites ($X^2_{4, N=287}$ = 20.456, p<.001): both groups featuring a high proportion of participants of European ancestry (69% in the laboratory, 85% online,) but I saw a higher proportion of East Asian participants in the laboratory sample (14% vs 4%). Screen resolution in the laboratory was 1920 x 1080, median screen resolution online was 1440 x 900.

The laboratory condition included ~27 participants in each task variant and the online condition included ~68 participants in each task variant (Table 3.2). Precise allocation of equal numbers of participants to each task variant could not be achieved online due to multiple concurrent signups to the experiment-platform.

3.4.2 Go Trial Reaction Times

Data from Go trials in all three variants and on both sites are shown in Figure 3.2, Figure 3.3 and Table 3.2. A two-way ANOVA of Go RTs indicated strong evidence for effects of both task variant ($F_{2,281}$=174.891, p<.001, η^2=.56), and site ($F_{1,281}$=24.906, p<.001, η^2=.08); however, there was no clear evidence of an interaction (p=.30). Go RTs were longer online and were also longest in the theme variant. Post-hoc t-tests showed that RTs from the theme variant were longer than the points (t_{190} = 16.316, p<.001, d=2.37) and non-game (t_{186} = 16.991, p<.001, d=2.49) variants; however, I could not detect a difference between the points and non-game variants (t_{192}=.085, p=.93, d=.01). A Bayesian t-test showed substantial evidence that Go RTs were equal in the non-game and points variants (BF=.16).

I also performed an exploratory analysis into the effect of task duration on RT (Appendix G - pg143), and found evidence that RTs shortened over the course of the task, but there was no evidence this change differed between task variants.

Table 3.2 Mean data from Go and No-Go trials, shown by site and task variant

			Go-Trials		No-go Trials	
Site	**Variant**	**N**	**Median RT (95% CI)**	**Accuracy (95% CI)**	**Median RT (95% CI)**	**Accuracy (95% CI)**
Online	Non-game	67	401ms (392 - 410)	97% (96.2 - 97.8)	368ms (350 - 385)	93.8% (92.5 - 95.1)
	Points	71	402ms (393 - 412)	98% (97.5 - 98.6)	372ms (355 - 389)	95.1% (94.2 - 96)
	Theme	65	481ms (473 - 490)	88.3% (86.1 - 90.5)	467ms (457 - 477)	65.7% (61.7 - 69.6)
Lab	Non-game	28	376ms (366 - 387)	99.4% (99.1 - 99.7)	356ms (336 - 376)	93% (90.7 - 95.3)
	Points	28	377ms (365 - 388)	99% (98.3 - 99.7)	351ms (331 - 370)	93.1% (91.1 - 95.2)
	Theme	28	469ms (459 - 480)	92.3% (90.6 - 94)	448ms (436 - 461)	65.6% (61.1 - 70.2)

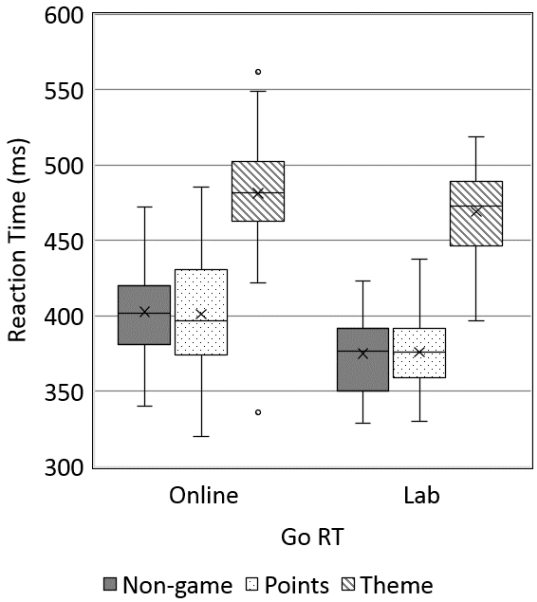


Figure 3.2 Box and Whisker plots of Median Go RTs, shown separately by task variant and test location

3.4.3 Go Trial Accuracy

Accuracy followed a similar pattern. A two-way ANOVA found evidence for main effects of both task variant ($F_{2,281}$=72.974, p<.001, η^2=.34) and site ($F_{1,281}$=15.277, p<.001, η^2=.05). Again, there was no evidence of an interaction (p=.14). Go accuracy was generally very high, as expected, although it was slightly lower online (Figure 3.3). Post-hoc t-tests showed that the theme variant had lower accuracy than the points (t_{190}=10.347, p<.001, d=2.03) and non-game (t_{186}=9.413, p<.001, d=1.75) variants. I saw no evidence for a difference between the points and non-game variants (t_{192}=1.511, p=.13, d=.23) and a Bayesian t-test provided insufficient evidence to support either equality or a difference between points and non-game variants

(BF=.46). Due to the non-normality of the data, I used Mann-Whitney U tests to confirm the ANOVA findings (Appendix H - pg144).

3.4.4 No-Go Trial Accuracy

Data from No-Go trials in all three variants and from both sites are shown in Figure 3.3 and Table 3.2. A two-way ANOVA of No-Go accuracy data found evidence for an effect of task variant ($F_{2,281}$=247.362, p<.001, η^2=.64), but no evidence for an effect of site or an interaction (ps>.393). No-Go accuracy was much lower in the theme variant than the other two variants, and post-hoc t-tests showed that the theme variant was different to the points (t_{190}=18.396, p<.001, d=3.57) and non-game (t_{186} = 17.582, p<.001, d=3.28) variants. Again, I saw no clear evidence of a difference between the points and non-game variants (t_{192}=1.012, p=.31, d=.15), but a Bayesian t-test found good evidence that No-Go accuracy was equivalent in the non-game and points variants (BF=.25).

I performed an exploratory analysis into the effect of task duration on No-Go accuracy (Appendix I - pg145) and found evidence that No-Go accuracy worsened over the course of the task, but there was no evidence this differed between task variants.

I saw ceiling effects in both the points and non-game variants, which resulted in skewed distributions. Due to the non-normality of the data, I used Mann-Whitney U tests to check the results of the post-hoc t-tests of Go and No-Go accuracy between task variants (Appendix H - pg144). All Mann-Whitney U tests confirmed the findings of the t-tests.

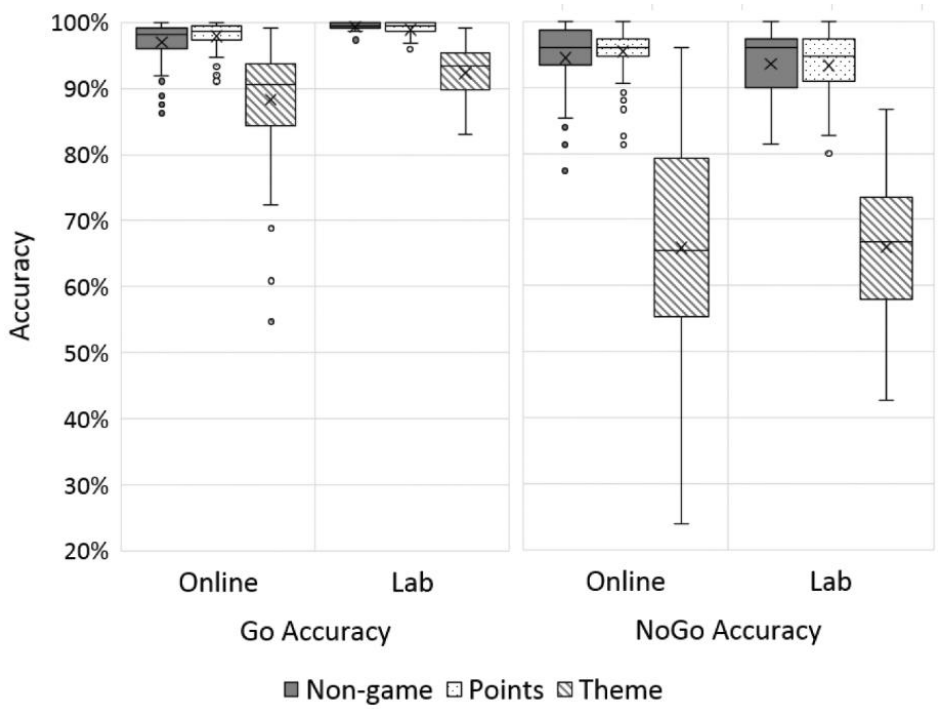


Figure 3.3 Box and Whisker plots of Go and No-Go accuracy, shown by task variant and site

3.4.5 Quality of Engagement

Figure 3.4 shows the mean scores from the assessment of quality of engagement, by site and task variant. A two-way ANOVA of total-score data found evidence of a main effect of task variant ($F_{2,281}$=3.719, p=.025, η^2=.026) and site ($F_{1,281}$=5.756, p=.017, η^2=.02), but no evidence of an interaction ($F_{2,281}$=.160, p=.85, η^2=.001). Quality of engagement scores were higher online (t_{285}=2.413, p<.016, d=.31), with online participants rating the tasks as more repetitive than those in the laboratory group, but being much more willing to take part in the study again. Post-hoc *t*-tests showed that the non-game variant provided a lower quality of engagement than the points variant (t_{192}=2.986, p=.003, d=.43), but no other differences were found (ps>.178). Bayesian *t*-tests were inconclusive as to whether quality of engagement was equal between the points and theme variants or the non-game and theme variants (BF=.32 and .37 respectively).

I suspected that heterogeneity in group composition might be responsible for the difference in quality of engagement between the laboratory group and the online group. Consequently, I performed a two-way ANCOVA of quality of engagement with task variant and test location as factors, and age and sex as covariates. I saw weak evidence for an effect of task variant ($F_{2,278}$=2.725, p=.067, η^2=.019), but not for site, sex, age or an interaction (ps>.103). This implies that the difference in scores between the two sites was indeed an artefact of age/sex preferences, and that task variant was driving a minor difference in scores (Figure 3.4).

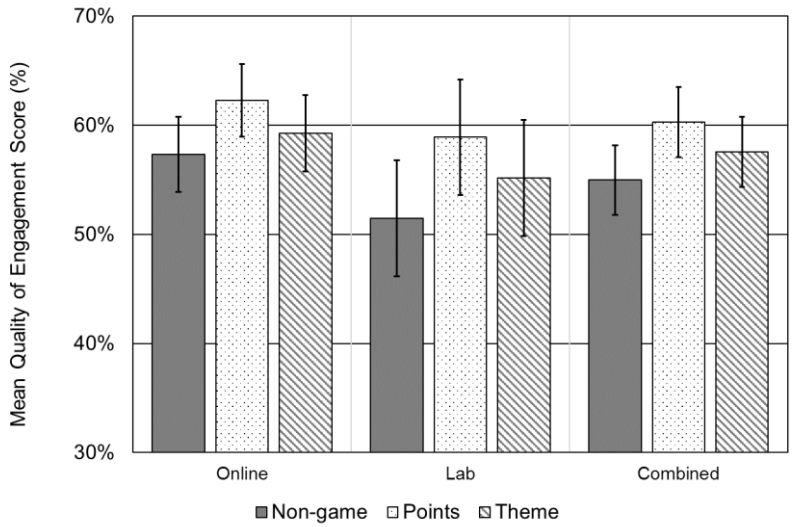


Figure 3.4 Mean total scores from the assessment of quality of engagement, shown separately by task variant. The combined score takes the average across both sites, after adjusting for age and sex. Error bars represent 95% CIs.

Figure 3.5 displays individual item scores, shown separately by task variant. The non-game control was rated least favourably on more than half of the items, including “boring” and

"frustrating". The theme variant had mixed scores, with participants feeling they performed poorly and finding it frustrating. However, it does appear that the cowboy stimuli resulted in the task being less repetitive, and on several measures, such as enjoyment, it does not differ from the points variant. Overall, the points variant was the best received: and participants who completed this variant were the most willing to recommend the study to a friend and reported putting the most effort into the task. I found no difference between any of the three variants on ratings of "difficulty concentrating" or "intuitive pictures".

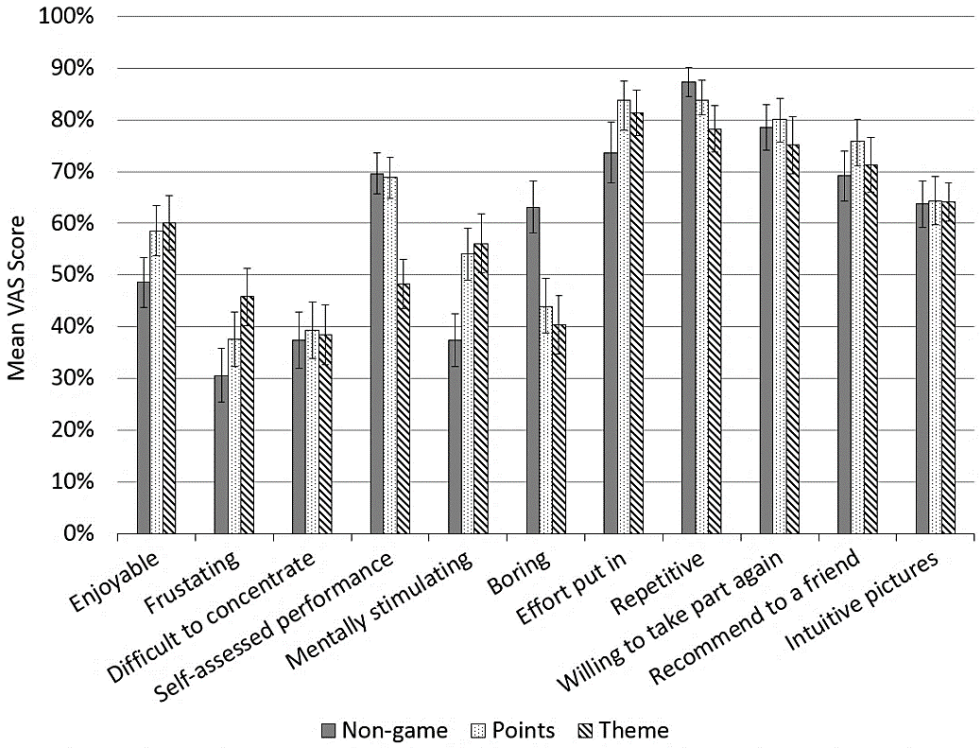


Figure 3.5 Scores for individual questions on the assessment of quality of engagement, shown separately by task variant. Error bars represent 95% CIs.

3.5 Discussion

3.5.1 Comparison of Task Site (Online vs Laboratory)

The laboratory group was unrepresentative of the general population in that it consisted mainly of young, female undergraduates who volunteered for the experiment. The MTurk group had a more balanced demographic, with a range of ages, education levels and games experience. Although MTurk users are also a self-selected group, their wider demographic lends ecological validity to the findings.

There was no evidence for a difference in No-Go accuracy between the sites, implying that online participants were just as able to inhibit their responses to stimuli. However, the median RTs of MTurk users were ~25 ms longer on average, potentially a result of lower participant

effort (due to the absence of an experimenter), environmental distractions or a difference in perceived reimbursement value. This stands in contrast to previous evidence suggesting MTurk participants have higher levels of attentiveness and performance than laboratory participants [188]. One might expect age to play a role in longer RTs, yet I found no evidence of a correlation between age and median Go RT, nor did I see an increase in No-Go accuracy in accordance with longer RTs, as would be expected [189,190]. This suggests that the difference may be due to technical reasons: although I used the same experimental platform to test both groups, there are still several potential sources of slowing such as differing operating systems, keyboards, web browsers and hardware [4,191–195].

Despite the difference in RTs, I saw no interactions between site and task variant, and there were no unusual patterns of performance between the two groups. These results show that online testing can produce valid and useful data if increased RTs are acceptable.

Against my expectations, the online group rated *all* task variants higher on the assessment of quality of engagement than the laboratory group. This is in contrast to Hawkins and colleagues [47] who reported lower engagement scores when the task was delivered online. Adjusting for age and sex eliminated the difference between the sites, implying that it was not the difference of testing location that influenced enjoyment, but rather the difference in sample composition. The theme variant was rated particularly highly online, and this may be due to the greater levels of gaming experience in the online group.

3.5.2 Comparing Task Variants

Contrary to my hypotheses, the data suggest the theme variant was far more difficult than the other two variants, with markedly longer RTs and lower accuracy rates. I propose several possible reasons for this: increased difficulty of spotting stimuli against the background image, a reluctance to shoot people (even cartoon characters) and the complexity of the stimuli. There was likely too much overlap in colour and pose between the civilians and cowboys, resulting in a slower categorisation of Go and No-Go stimuli.

My motivation for using red and green objects as opposed to basic stop and go symbols was to match the intuitiveness of stop and go stimuli across task variants (i.e., I felt that shooting the cowboys and avoiding the innocents would be so intuitive that the non-game variant would need equally intuitive stimuli). However, the association between red/green and stop/go may have been stronger than I expected [180] and there is evidence that attending to colour is easier than attending to shape [196]. These factors may have made the points and non-game variants easier than anticipated. Any implicit association between red/stop and green/go may

have been unnoticed by participants as they reported that stimuli in the theme variant were equally intuitive to those in the points and non-game variants.

The differences between the theme and non-game variants demonstrate that using complex gamified stimuli in cognitive tasks can be problematic. This is an important finding since several previous studies have used complex stimuli, such as robots and monsters, in their gamified cognitive tasks [52,92,197]. The idea of using graphics alone to gamify a task is not uncommon, but future researchers must ensure that the addition of gamelike stimuli does not make their task more difficult. Detrimental effects on participant performance resulting from the introduction of game design elements have been found before [49,101], and it is likely that complicating a task too much may increase its difficulty such that the data it collects becomes incomparable to data from a traditional task.

That said, Boendermaker and colleagues [72] investigated the use of game elements in a GNG alcohol-bias training task. Their game variant was themed and contained points, lives and levels, but they saw no difference in training efficacy between the gamified and non-gamified variants. Their results stand in contrast to my minimally themed variant which had a negative impact on participant performance. This may be because Boendermaker's task clearly delineated the task stimuli from the themed surroundings of the game (i.e., using extrinsic fantasy) [76], rather than redesigning the actual stimuli as I did.

When I consider the data collected by the non-game and points variants, Bayesian *t*-tests provided good evidence that these task variants produced equivalent data. The points system was not particularly punishing and this may explain why I saw no impact of the points system on behaviour. There is evidence that a GNG task which rewards participants for fast responding and punishes them for failed inhibitions can optimise performance [181]. However, despite predictions that gamification might boost participant performance [47], I did not detect any improvement in data as a result of the points mechanic. The points variant received the highest total engagement score of the three variants, both online and in the lab. This is interesting because adding points to cognitive tasks to make them more engaging is not uncommon, but to the best of my knowledge this is the first study to directly compare the appeal of points against another game element.

Finally, it is clear from the results that the addition of even a single game design element can make a difference to participants' perception of a task. In line with my hypothesis, the non-game variant was rated as more boring, less enjoyable and less mentally stimulating than either of the gamified task variants. The results show the theme variant to be of secondary

appeal to points, but this may have been influenced by the fact that the theme variant was more difficult. As such, it comes as no surprise that participants rated it as more frustrating and felt they performed less well. Future work might investigate the role of theme more effectively by carefully controlling task difficulty. I also highlight the need for replication of these findings, with points being compared against other themes or in other contexts, such as longitudinal studies.

3.5.3 Limitations

I consider the difference in difficulty between the theme variant and the other task variants to be the most important limitation of this study. This difference is informative because gamelike stimuli and complex visual environments are common in gamified tasks, and the results highlight the need to limit the impact of these features. The resultant variations in accuracy clearly limit the extent to which one can compare task performance across variants. Secondly, I opted for a between-subjects design: which did not allow me to study the impact of different game design elements on an individual's performance and confounds hardware/individual differences with effects caused by the task variant. Nevertheless, the large sample size I achieved using online testing helps to counteract the lack of power associated with this experimental design. I also acknowledge that my design is not suitable to validate the task for the measurement of response inhibition, and that I would require a within-subjects design in order to test predictive validity [152,198]. Thirdly, the task I used was short, and participants may not have had time to become bored enough to affect the data, even when playing the non-game variant. If participants were not bored by the task, this would have limited any motivational effects of gamification. Future research might explore whether longer task durations result in greater boredom, and therefore greater impact of gamification. Although I intended the questionnaire to capture participants' quality of engagement, the fact that it was delivered after the task means it represents only a post-hoc appraisal of the task. Although this is a common method for assessing quality of engagement (see Section 2.4.5), it weakens the measure, since a participant's memory of how engaging the task was could be quite different to their experience at the time. Finally, based on the definition of engagement in Section 2.5.1, I only assessed the effect of gamification on one dimension of the concept of engagement. This study lacked a measure of amount of engagement, thus limiting the generalisability and impact of these findings.

3.6 Chapter Summary

In this chapter I described an empirical study into the effect of adding individual game elements to a cognitive test. I used the GNG task, a common cognitive test which measures

response inhibition, and created three task variants: a non-game control, a variant where participants were awarded with points in line with performance, and a graphically themed variant framed as a cowboy shootout. I tested participants across two sites: in the laboratory and online to compare the suitability of online testing for further exploration of this field. The result was a 2 × 3 between subjects design, with task variant (non-game, points, theme) test location (laboratory, online) as factors.

I found points to be a promising game design element for gamified cognitive testing: they did not disrupt the validity of the data collected and increase quality of engagement. However, despite some hope that game design elements might increase engagement to the point where participant performance improves, I found no evidence of such an effect in this study.

The data show that while participants enjoyed the themed task and its visually interesting stimuli, the complexity of categorising such stimuli detrimentally affected participant performance. The lowered accuracy rates and increased RTs I saw were likely the result of the increased visual complexity of the stimuli. In hindsight, cowboys and innocents were simply much harder to differentiate than red and green objects. Future studies should match stimuli across gamified variants more carefully, with the human visual system in mind.

I saw differences in the data collected online and in the laboratory, with slightly longer RTs in the online group. However, there were no interactions or unusual patterns of performance, suggesting that online crowdsourcing is an acceptable method of data collection for this type of research.

Overall, I set out to establish a methodology for further investigating the effects of gamification on cognitive data and engagement, and I achieved that goal. I employed individual game elements (points and theme) and demonstrated that they had distinct effects on performance and quality of engagement. I showed that online testing was a suitable method of data collection, and that it produced a pattern of results comparable to the laboratory.

There is still considerable scope for improvement. Firstly, the poorly matched stimuli in the theme variant could be designed to impact less on the difficulty of the task. Secondly, I did not measure amount of engagement, thus capturing only part of the multidimensional definition of engagement. Thirdly, the task was hosted and developed by a third party, Xperiment.mobi. Though the task was programmed according to a specification I created, and the design of the gamified variants was iterated between myself and the programmer, the indirect nature of my

involvement meant I that struggled to achieve my desired level of creative control. In the next chapter, I document the creation of a bespoke platform for delivering gamified cognitive tasks online. In Chapter 5, I describe the use of this platform to run Experiment 2, addressing the other limitations discussed above.

Chapter 4: The Mindgames platform

The code for the Mindgames platform is archived on Zenodo, doi: 10.5281/zenodo.1477612

4.1 Chapter Aims

This chapter documents the development of Mindgames: a custom web app developed for the delivery of gamified cognitive tasks. This chapter does not describe an empirical study, rather it is a retrospective examination of the design decisions I made during the development process. The reader might consider reading this chapter in parallel with Chapters 5 and 6, to have more context on how the Mindgames platform was used. This chapter has three aims, namely to:

1. Specify the requirements of the platform
2. Describe how those requirements were met
3. Evaluate the platform

4.2 Introduction

In the previous chapter I described my first empirical study using a gamified cognitive task. This task was developed by an external contractor and hosted on the Xperiment platform. The same contractor also managed participant recruitment, support and payment. This method had advantages (primarily convenience) and several disadvantages. Firstly, I wanted to run a longitudinal study, which would require functionality Xperiment could not provide. Secondly, I sought more creative control over the task in order to deliver richer gamification (animations, complex graphics and menus, etc). Thirdly, I wanted to move from an app-engine optimised to run cognitive tasks (Xperiment) to one optimised to run browser-based games, with the aim of minimising performance disparities between users on different hardware, and enabling new functionality.

At the time this project began, there were limited available tools for running longitudinal online studies and so I needed to develop my own. However, the last few years have seen the emergence of several websites for creating and hosting complex online cognitive tasks, including Gorilla, Testable and the recently launched Xperiment 2.0. Known as ‘platforms’, they allow researchers to construct their own tasks using scripting and graphical user interfaces. They are capable of hosting and delivering a variety of cognitive tasks and provide a raft of supporting technologies (user identification, data security, storage, scalability, etc.).

Development of my own platform, Mindgames, was iterative. I used it to run Experiments 2 and 3 in addition to three further studies which I collaborated on during my PhD [199–201]. After each study I assessed the platform’s strengths and weaknesses and made appropriate modifications. In this chapter I present Mindgames as it was during the final study of my

thesis; focussing on the design process of the technology rather than the specifics of cognitive tasks deployed on it.

The code for Mindgames used to run Experiment 2 can be accessed here:
github.com/jl9937/longitudinalGamifiedStopSignalTask
The code for Mindgames used to run Experiment 3 can be accessed here:
github.com/jl9937/SingleSessionBoredomStudy
A demo of Mindgames as it was during Experiment 3 can be found here:
mindgamesmkii.firebaseapp.com

4.2.1 Using Prolific

In Experiment 1, Amazon MTurk was used to recruit participants, however given I was changing study platform, I also had the opportunity to change recruitment website. At the time of launching Experiment 2, there were two main choices for online participant recruitment: MTurk and Prolific.

These two online recruitment services have several differences. MTurk is well established and claims a user-base of 500,000 workers [202]. MTurk has a broad scope, hosting jobs in image/video processing, and data verification and collection. Most of MTurk's participants are based in the US and in India [203]. MTurk has no minimum participant payment per hour [204], and is quite simple in its organisation: one cannot easily screen users to meet certain criteria. Importantly, paying participants on MTurk requires a US bank account.

In comparison, Prolific is newer and claims a user base of only 25,000, most of which are from the UK and the US. Prolific focusses on providing participants for academic research studies and allows the screening of participants to meet specific requirements [205]. Submissions from respondents can be reviewed and approved before payment, and an hourly rate of £5 minimum per hour is enforced.

Both MTurk and Prolific have been used to recruit participants for online behavioural studies, and results from both platforms are comparable, in addition to replicating laboratory findings [1,3,206]. However, some have raised concerns over a lack of participant naivety on MTurk [207], where some participants are "professional survey-takers" [208]. Furthermore, there is evidence that the active participant pool on MTurk is orders of magnitude smaller than 500,000: perhaps as small as 7000 [202]. In contrast, participants on Prolific are more naïve [3], and the active population is in the region of 20,000.

Given that the active participant pools of MTurk and Prolific are similar in size, that Prolific does not require a US bank account, and that participants on Prolific have been shown to produce high quality data: I decided to use Prolific as a recruitment service. I therefore designed Mindgames specifically to interface with Prolific's website.

4.3 Platform requirements

Mindgames was required to:

1) Support anonymous, signup free participation

To keep signup as hassle-free as possible, users needed to be able to simply follow a link and begin the study immediately. Signup needed to be anonymous, as collecting participant email addresses violates the terms of service of Prolific.

2) Support longitudinal designs

Mindgames needed to be able to recognise returning participants, load their profile data and track their progress through sessions of a study. The platform also needed to be able to deliver different session procedures to different participants on different days, depending on their progress. Finally, the platform needed to ensure adherence to the study protocol, for example, limiting participants to completing only one session per day if required.

3) Be game suitable

The platform needed to run tasks with game elements. It needed animation support, image caching, low memory usage, responsive input and a fully customisable appearance.

4) Scalable

Online psychological tests have the advantage of assessing multiple participants simultaneously, but doing so requires a webserver fast enough to handle many connections at once and a database capable of accessing multiple records concurrently. The platform needed a webserver and database that would be able to serve multiple participants at once, and scale up when demand was high, ensuring a consistent experience for all users.

5) Secure

According to the Data Protection Act 1998 [209] I, as a data manager, was required to store personal information "safely and securely". Cognitive measures, certain demographic information and identification numbers all constitute personal information. Mindgames' database therefore needed to be secured against unauthorised access, despite being part of a login-free system. I also required that certain types of data were write-only: inaccessible to anyone other than study administrators.

6) Provide precise response timing

Many cognitive tests derive their output measures from RT. Accordingly, Mindgames needed to record response times with millisecond precision.

7) Provide accurate stimulus presentation timing

Stimulus presentation is important to cognitive tasks for a range of reasons: to ensure consistency in display, to prevent conscious study, to encourage rapid response, etc. Accordingly, Mindgames needed to ensure accurate stimulus presentation times.

8) Minimise cross platform differences

Online tasks will be accessed on a range of hardware (from smartphones to PCs) and through a range of web browsers (from Internet Explorer 10 to Safari). Varying levels of cross-browser support for language features make it difficult to deliver a consistent user experience. Nevertheless, Mindgames needed to deliver as consistent a participant experience as possible to ensure the comparability of data between users.

9) Extensible and Configurable

Mindgames needed to be easily modifiable to support different types of studies. Properly designed, it should be easy to alter the platform to deliver different cognitive tasks, new study procedures, new questionnaire types, etc.

10) Provide management utilities

Mindgames would be used to run studies with hundreds of participants. These participants could be at different stages of study completion, they might be missing data, have experienced bugs, etc. Mindgames needed to provide tools to monitor participants, manipulate the database and support the downloading of large data files.

4.4 Implementation

Mindgames was designed as a single page webapp, written in Javascript and HTML 5. I used Visual Studio 11 as an IDE and NodeJS to facilitate website deployment. For brevity, I do not provide detailed design documentation on all of Mindgames. Instead I discuss four core design decisions/technical solutions which addressed the majority of platform requirements.

4.4.1 Core Libraries

Firebase

I solved several web-hosting, database storage and security problems at once by using Google Firebase. Firebase provides an access library and lightweight JSON-like database (JavaScript

Object Notation), storing data in a tree-like structure of objects and properties as opposed to a more traditional tabular-record format (Figure 4.1). The resemblance of Firebase's data storage structures to JavaScript objects makes the two extremely easy to integrate: one can write JavaScript objects to the database in a single line and read them just as easily. Importantly, Firebase provides cloud-based web hosting and database storage, providing scalability to hundreds of instances in addition to robust backups. The cost of this service depends on usage, but I found it to be very low: less than £10 to run five longitudinal studies with over one thousand participants in total. Firebase's web hosting service is not sophisticated, allowing only very limited server configuration. Nevertheless, given the richness of the rest of the Firebase library, I decided that building the webapp without any server-side computation was a limitation worth accepting.

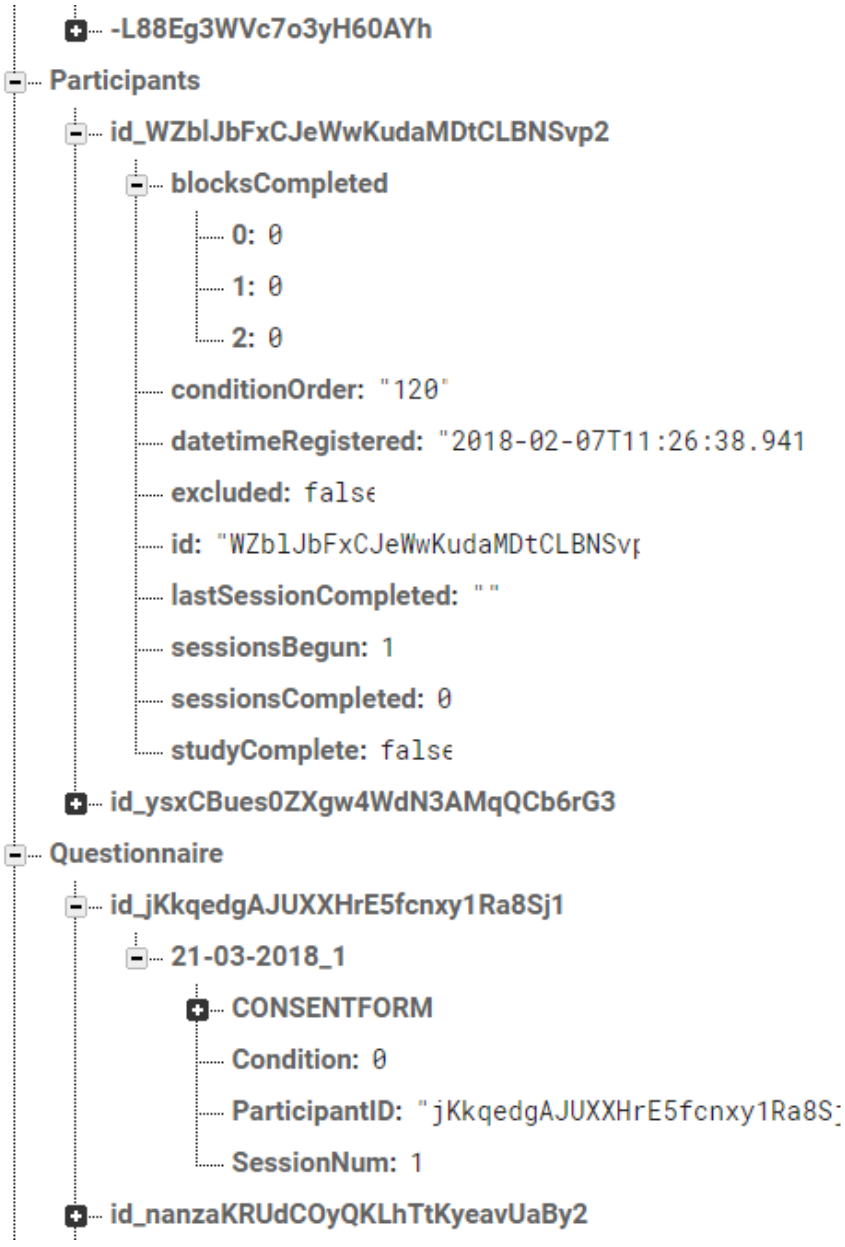


Figure 4.1 Screenshot from the Firebase database interface showing the treelike structure of data. In this image, the Participants node contains data on two participants. One participant node is expanded to show the data stored on that participant.

Firebase's data storage functionality is supplemented by strong security features. Security rules enforce database structure and can restrict access to nodes of the tree. For example, I specified that a participant's client (the Mindgames code running in their browser) could only read/write to objects with an ID which matched its own user ID (UID) (Figure 4.2). These UIDs were uniquely associated with a username-password pair, making reading or writing another participant's data impossible unless their credentials were guessed. Wherever possible, I took a write-only approach to data storage, with most security rules specifying that only one UID (the system administrators) could read the data stored at that node and its children. RTs, questionnaire responses and user navigation activity were all stored using this one-way method. The participant's client could only read data essential to the running of the webapp, e.g., the experimental condition the participant was assigned to, the total financial incentive they'd accumulated, how many sessions they'd completed, etc. The combination of Firebase's security rules, cloud-based hosting, database service and library played a large role in helping me meet the requirements of scalability and security.

```
130 ▾    "Participants": {
131 ▾      ".indexOn": [ "excluded", "sessionsCompleted" ],
132 ▾      "$id": {
133          ".read": "$id.endsWith(auth.uid)",
134          ".write": "$id.endsWith(auth.uid)",
135 ▾        "conditionOrder": {
136            ".write": "!data.exists() || !newData.exists()"
137          },
138 ▾        "blocksCompleted": {
139 ▾          "0": {".validate": "newData.isNumber() && newData.val() >= 0 && newData.val() <= 20"},
```

Figure 4.2 Screenshot of the Firebase security rules interface. The lines highlighted in yellow show that read and write access to a Participant node is only granted when the node's name matches the user's authorisation UID.

PixiJS

I also made extensive use of PixiJS: an open-source, Javascript based, 2D rendering engine built by Goodboy Games. The PixiJS library provides functionality for creating 2D graphics in JavaScript: for example, it includes a suite of classes to support sprites and their placement on an HTML 5 canvas. These classes handle the caching of textures (so they are downloaded before they're needed), the tweening of animations (frame by frame smooth blending of animations) and maintain a structure of parent and child objects (allowing composite graphical elements to be created). PixiJS manages the rendering loop and provides cross-browser optimisation to keep the webapp's framerate as high as possible, across a range of devices and browsers.

Due to the complexity of PixiJS's key functionality (the rendering loop), the library is quite low-level. I had to extend several PixiJS classes to provide requisite functionality such as configurable buttons and questionnaire input elements (VAS, number-pickers, etc). Nevertheless, the use of PixiJS was core to making the platform 'game suitable' and minimising cross-browser differences.

4.4.2 Anonymous Login Procedure

The Problem

I required the login process to be anonymous, secure and invisible to participants. There were several challenges to implementing this functionality: firstly, Firebase provides an authenticated login service, whereby a user can login with a valid username-password pair and receive an authentication token. This token holds two important pieces of information: the auth-key and the UID. The UID is a 24-character alphanumeric string which identifies the user, and the auth-key is a 64-character secret-key which authenticates every message to Firebase. This login service was vital to ensuring the security of data stored in the platform; but was not immediately compatible with my requirement for signup free participation due to the need for a username-password pair.

Secondly, users of Prolific are identified by their unique ProlificID. This takes the form of a 24-character alphanumeric string which the participant must submit to the study-platform to prove that they have tried to take part. Not every participant who tries to take part will manage to complete the study, with some dropping out due to technical difficulties, interruption, etc. At the end of the study, the researcher compares their platform's list of ProlificIDs that completed the study against Prolific's list of ProlificIDs that signed up to take part. They can then reward participants appropriately. However, ProlificIDs are not anonymous. The same ProlificID is used by the participant in every study they complete, making it possible to identify an individual if data is collated across many studies. ProlificIDs therefore constitute personal information and must be secured.

Thirdly, I required that the platform was suitable for running longitudinal studies. I needed to be able to recognise returning users and load the appropriate settings into the study page: potentially moving them onto the next session, allowing them to continue from where they left off, etc. Given my intention to use email-password pairs, I needed a method of remembering that information between sessions, without requiring the participant themselves to remember it.

I therefore needed to build a login/signup system that could generate an email-password pair, remember login details across sessions and record ProlificIDs while preventing data-linkage between those IDs and other personal information.

The Solution

When a participant signs up to take part in a study on Prolific, they are given a link to the study-platform (mindgames.firebaseapp.com/task.html). I used an autofilled URL parameter to append the participant's ProlificID to the end of this link so that when they loaded the study page, I could programmatically extract their ProlificID from the URL. I split the 24-character ProlificID into two halves, using the first twelve digits prepended to "@mindgames.com" as a false email address, and the last 12 digits as a password. The result was a deterministic, autogenerated email-password pair, which could be used to log participants in automatically as they loaded the platform. Finally, as mentioned above, Firebase's login procedure provides an authentication token with a unique UID that is not derivable from the email-password pair. I used this UID as the internal identifier under which all data associated with that participant was stored: thus ensuring that the data were pseudonymised in the event of a security breach. I used a write-only node in the database to store ProlificID-UID pairs so that I was able to cross-reference signups on Prolific with participants in my study, but this node was only readable through the admin interface of Google Firebase, which was secured behind a strong password and 2-factor authentication.

This implementation of the login procedure met my requirements to be signup free, anonymous, secure and to support longitudinal designs without requiring participants to remember login details.

4.4.3 Accurate Stimulus Presentation Times

The Problem

Ensuring accurate stimulus presentation times was a challenge due to a combination of factors: firstly, JavaScript runs in a sandbox, meaning that the speed of its execution is dependent on the resources granted to it by the host operating system. Secondly, JavaScript uses automatic memory management, which periodically pauses execution to delete objects that are no longer needed. Thirdly, JavaScript timers are asynchronous and notoriously inaccurate. Fourthly, PixiJS makes use of a rendering loop (mainloop), making it difficult to synchronise animation frames with timers.

PixiJS's mainloop repeats continuously as the task runs, reacting to user interactions, advancing the task state, storing data, etc. At the end of each loop, the program recalculates

what should be shown on the user's screen and draws a new frame of the animation; hence the execution time of the mainloop determines the task's framerate. The execution time of the mainloop is variable, with some iterations requiring more processing than others (i.e., a loop where a trial starts takes more time than a loop where the task simply waits for user input). Furthermore, when automatic memory management or resource-constraint occurs, the execution of the mainloop is momentarily paused resulting in a 'slow-loop' and a temporary drop in framerate. Smooth animation requires a framerate of at least 40 frames per second, meaning the mainloop must execute in under 25ms.

The conventional JavaScript approach to displaying a stimulus for 100ms would be to use the built-in, threaded function *SetTimer.* A threaded function runs independently (asynchronously) from the mainloop: it performs its task (waiting 100ms), then updates the application state (removes the stimulus) and is destroyed. However, there are two problems with using *SetTimer* to display stimuli: firstly, the content of the screen cannot change until the mainloop ends and the frame is redrawn, thus, although *SetTimer* is asynchronous, its precision is bounded by the framerate. Secondly, *SetTimer* is not designed to guarantee accuracy, and frequently overruns its timer.

In initial testing, the combination of inaccurate *SetTimers,* poorly synchronised frames and unpredictable mainloop-execution times caused serious difficulties for accurate stimulus presentation. Though the cooccurrence of slow-loops and inaccurate timers was unlikely, a 10-minute cognitive task involves 24000 iterations of the mainloop and 1280 calls to *SetTimer*, making the comparatively rare event occur many times during the task. The result was that some stimuli were displayed for less time than intended, while others lingered on the screen. It was clear I needed to implement a more accurate version of *SetTimer:* synchronised with task framerate and accounting for slow-loops.

The Solution

I abandoned *SetTimer* and wrote a new *Timer* class (Figure 4.3). I synchronised *Timer*s with the task framerate by moving them inside the mainloop rather than having them execute as threads. Furthermore, *Timer*s did not begin timing until a frame had been drawn.

On each iteration of the mainloop, Mindgames checked its internal list of *Timers*. Those which were not yet running were started (with a value of 0), and *Timer*s which were already running were incremented by the number of milliseconds which had passed since the last frame. After incrementing the *Timer*'s value, the mainloop then checked to see if the *Timer* was over its intended duration: if it was, its callback fired and the timer was destroyed.

While JavaScript's *SetTimer* (inaccurately) waits a certain amount of time before executing, my *Timer* repeatedly checks how much time has passed since it was started using *performance.now()*, and if it is over the target amount, it executes. *performance.now()* which provides the 'time since page load' in milliseconds, accurate to 5 microseconds. Importantly, the timestamp drawn from hardware and not based on JavaScript's *SetTimer* technology, and is therefore not affected by resource limitations.

My implementation of a timer system made use of hardware-accurate timestamps and was synchronised with the task framerate. This greatly increased the accuracy of stimulus presentation times and helped to minimise cross platform differences in performance.

```
function Timer(time, callback)
{
    this.targetTime = time; //number of milliseconds the timer should run for
    this.duration = 0; //number of  milliseconds the timer has been running for
    this.callback = callback; //function to be called when the timer completes
    this.alive = true; //is the timer currently timing?
    this.frameStarted = loopCounter; //the framenumber the timer began on
}

Timer.prototype.check = function(millisecondsSinceLastFrame)
{
    if (this.alive)
    {
        //ensure the frame has been displayed for at least 1 frame before we start timing.
        if (loopCounter - this.frameStarted <= 1)
            this.duration = 0;
        else
            this.duration += millisecondsSinceLastFrame;

        if (this.duration >= this.targetTime)
        {
            this.alive = false;
            this.callback();
        }
    }
}
```

Figure 4.3 JavaScript code describing the Timer class. A custom helper class used to increase the accuracy of stimulus presentation times.

4.4.4 Future Proofing

The Problem

My intention was that Mindgames could be used by future researchers to deploy a range of online cognitive studies. To achieve this goal, I needed to ensure that the platform's codebase was well structured, suitable for continued development, and that the platform provided management tools that would allow non-technical users to run studies.

The Solution: Object-Oriented Design

Mindgames was built using object-oriented design principles. i.e., the program was broken down into a series of self-contained building blocks. Object-oriented design has four main principles: single logical function, interfacing, inheritance and abstraction.

The principle of single logical function maintains that each building block, or *class*, is responsible for one logical area of functionality. The data needed to provide this functionality and the code which describes it, is contained within the class itself. There are several advantages to this design approach: firstly, functionality is logically portioned making it easier for a developer to find the code they're looking for. For example, if a new developer was searching for the code which handles stimulus presentation times, then the *Timer* class would be an obvious place to start.

The principle of interfacing maintains that classes should not directly access data stored in other classes: rather, the two classes should communicate via a function interface (Figure 4.5). This means that while the internal functioning of a class might change over the course of development, the external 'interface' of that class remains the same, and the rest of the webapp does not have to be rewritten to accommodate it. e.g., the specifics of how Mindgames stored data in Firebase changed several times over the years, but because the *DBInterface c*lass functions remained consistent, no modifications to *Session*, *Participant* or other higher-level classes were required. From the perspective of those classes, nothing had changed.

The principle of inheritance allows the programmer to describe relationships between classes (Figure 4.4). For example, the *View* class stores data and functionality common to all task screens, but a series of individual classes for each screen (*GenericScreen*, *Engine*, *ConsentForm,* MainMenu, etc) inherit *View* and contain only code and data specific to their role. Importantly, classes which inherit *View* can be interpreted by the rest of the program as either their specific instantiation (such as *ConsentForm*) or as a *View*. This simplifies programming as it allows other classes to interact with all the *Views* in the same way, regardless of their specific type.

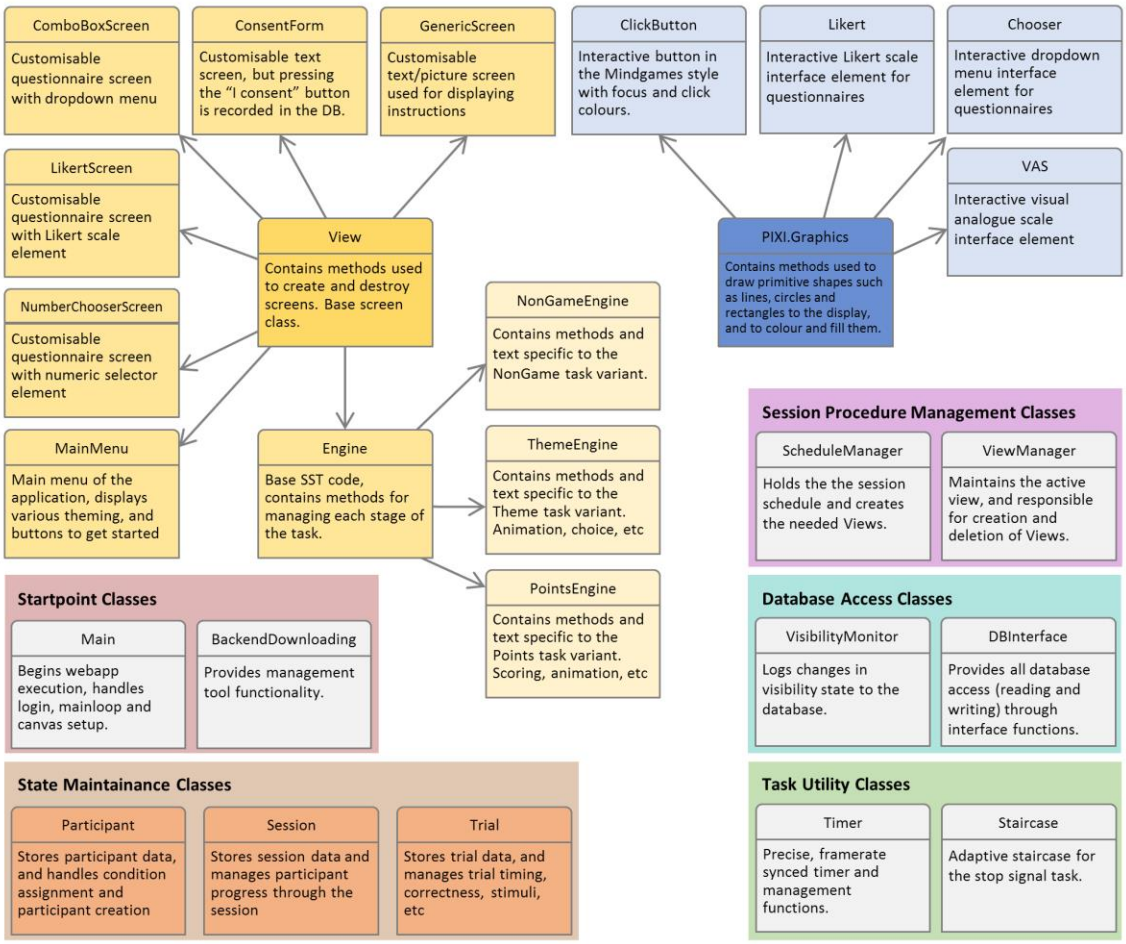


Figure 4.4 Inheritance diagram of classes in Mindgames. Classes with related functionality are grouped, and arrows show inheritance. The description below each class summarises key responsibilities and purpose.

The principle of abstraction maintains that the programmer should hide all unnecessary detail in the code from other objects to reduce complexity and increase readability and efficiency. Fourthly, big libraries such as PixiJS and Firebase offer a vast amount of functionality, meaning they can be applied to a range of projects. In the case of a platform like Mindgames the required functionality is narrow, and so much of the complexity of these libraries can be hidden. Using the principle of abstraction, I built wrapper classes that took PixiJS or Javascript functionality and gave it a simpler interface. For example, Mindgames' buttons are made up of rounded-rectangle Pixi.Graphics (with specified colour, position, corner radius, etc), that change colour when clicked. Click detection is done using a PixiHitArea (with specified coordinates and callbacks), and the text is created using PixiText (with specified font, position, size, colour, etc). Since all the buttons in Mindgames look the same, all these properties can be hardcoded and abstracted away into the *ClickButton* class. In the final iteration of Mindgames, a single line of code specifying a button's text and its callback function could add a button to the screen.

The goal of applying these design principles was to make the Mindgames codebase easy to understand and to simplify the development of other tasks on the platform. Figure 4.5 shows an example communication flowchart of the platform in operation. The central box (*NonGameEngine*) represents the sole class responsible for running the stop-signal task (SST), highlighting how much functionality Mindgames provides besides the cognitive task itself. Given the modular nature of the design, making changes to the SST or swapping it for a new task should be simple. Taken together, the design principles of single logical function, interfacing, inheritance and abstraction met my requirements to design a platform which would be configurable and extensible by future developers.

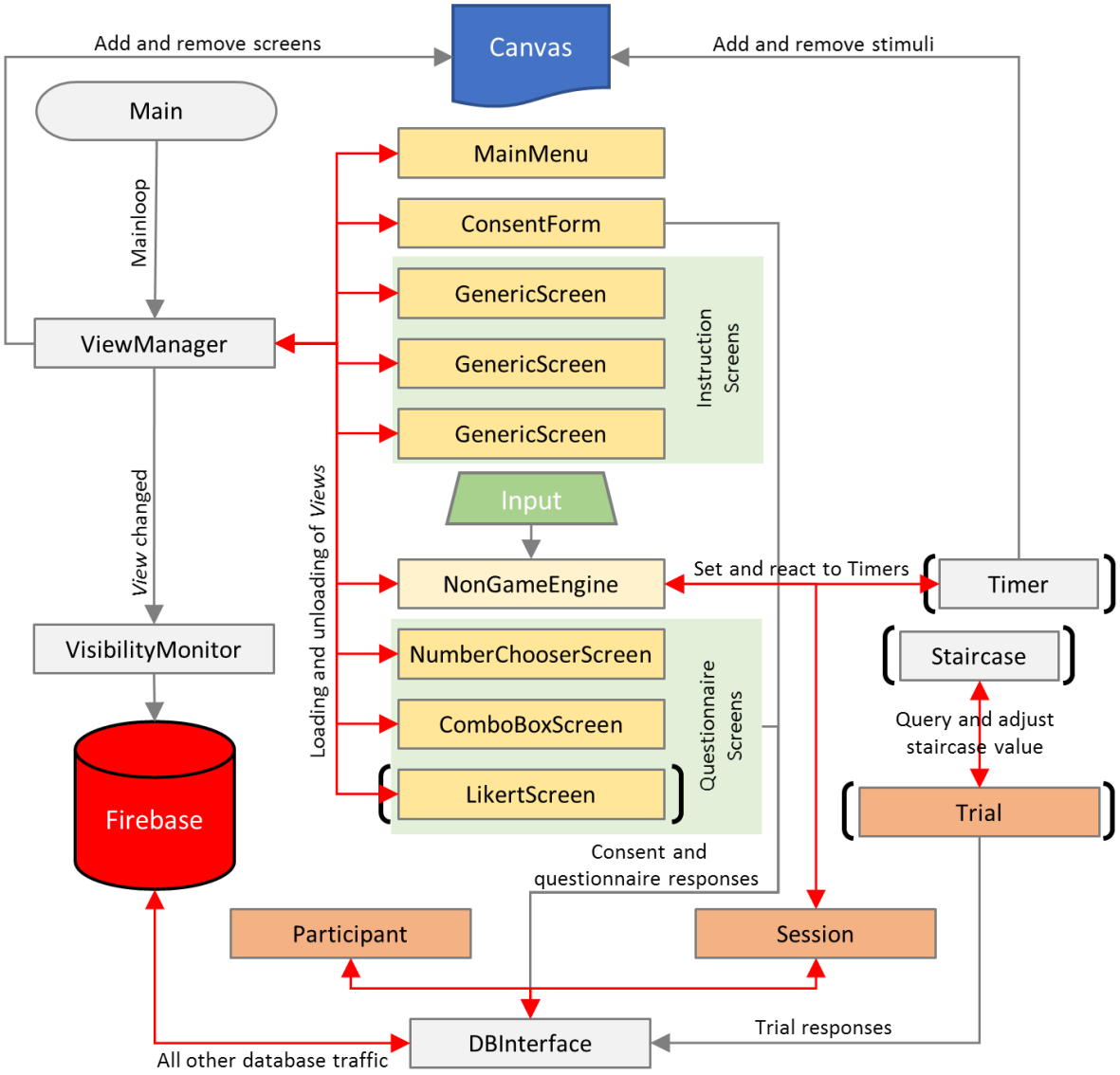


Figure 4.5 Example communications flowchart of Mindgames. This diagram depicts the flow of data to and from objects during the delivery of the non-game variant of the SST. The two green boxes denote sets of Views*: the instructions screens and the questionnaire screens. Red arrows represent two-way communication. Grey arrows represent unidirectional communication.*

Brackets denote arrays of objects. There are three endpoints of the system: the user's screen (in blue), the user's keyboard (in green), and the Firebase database (in red).

The Solution, part 2: Management Tools

In order that non-technical users could run studies on Mindgames, I needed to provide a suite of management tools for the database. The Firebase database can be manually accessed and modified through the Firebase interface (Figure 4.1), but this method of access has no safeguards. The user can modify any entry, delete any node, etc, and a single misclick could result in the deletion of the entire database. To provide database access with less room for accident, I created a second (smaller) webapp with a simple graphical user interface (Figure 4.6). This webapp allowed the user to download data, check the status of participants, exclude and unexclude participants and search the database for participants at a certain study stage.

Building on my principles of flexible design, I built download tools which were data structure agnostic. Thus, regardless of the tree-structure of the database, the downloader would automatically flatten the tree, handle missing data gracefully, tidy up variable names for display and download the database into a tabulated format.

The need for a means to flag participants as 'excluded' arose in early pilots: some participants load the Mindgames webpage, resulting in a profile being created for them, but do not go on to complete the experiment; some drop out of the study partway; some are temporarily accounts used to check for bugs or test certain features. These invalid participants clog up data downloads and status queries, making it difficult to monitor the progress of legitimate participants. I developed the exclude feature so that these participants could be hidden from data downloads and in the interface unless requested.

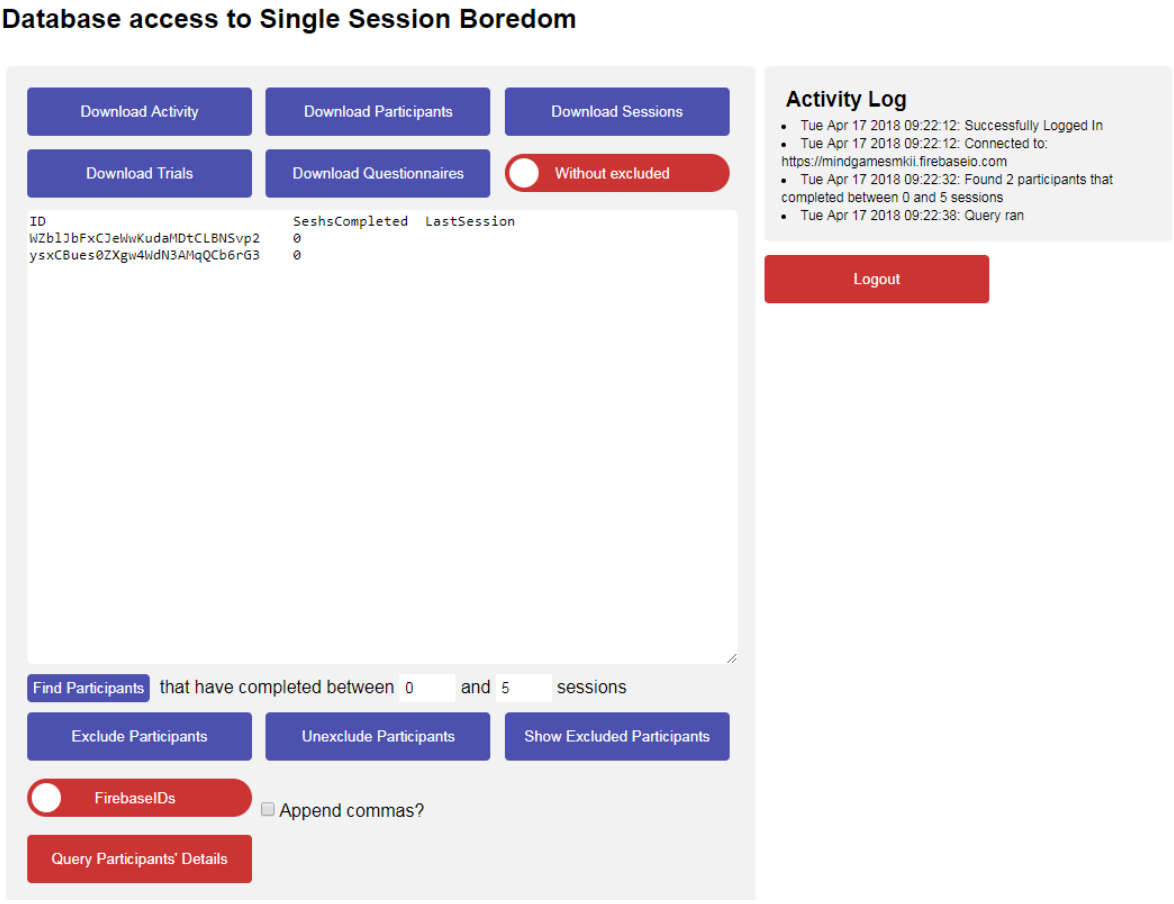


Figure 4.6 Screenshot of the Database access user interface, providing data-download functionality and safe database modifications tools.

4.5 Evaluation

Having used Mindgames to successfully run five large-scale online studies, I am satisfied with its functionality. However, as with any long-term project, there is room for improvement.

Mindgames' ability to run longitudinal studies is acceptable, though a little clunky. The *ScheduleManager* class defines the procedure of each session and creates the required *Views*, but this is hard coded and therefore not modifiable without some technical knowledge. A logical step for development would be to allow session procedures to be defined in separate script files (using non-technical language), which could be read into Mindgames as the site loads. Similarly, the operation of the task itself could be separated out into a script file, thus making the platform more task agnostic (i.e., more extensible and configurable). However, in the case of this thesis, the gamified task variants required substantial technical implementation and it would not be possible to extract them from the codebase.

PixiJS's library supported my gamification requirements and laid the groundwork for a mainloop driven app. This, in combination with my *Timer* implementation and *performance.now()* allowed me to ensure consistent response timing and stimulus

presentation timing across a wide range of devices. Timing was not perfect however and was still dependant on the user's computer. I had limited access to test on multiple devices but found that lower specification hardware had longer loading times and more significant delays when a slow-loop occurred.

Additionally, RTs recorded by Mindgames are longer on average than RTs from comparable non-JavaScript based tasks. Research suggests that timer latency in online tasks is not unusual (likely arising as a consequence of the sandboxed browser environment), and doesn't have an enormous impact on cognitive data collection [193–195]. Future work should involve experimentation to test the accuracy of recording RT in Mindgames, possibly using an automated test rig to generate mouse input with known response times.

Firebase's cloud hosting service allowed my platform to be scalable. This worked well, with Mindgames serving hundreds of participants simultaneously at peak times. Over the five studies it delivered, there was only one service outage and it did not seriously impact participants. Overall, my experience of Firebase was positive: their library was fully featured and the user interface was powerful and navigable. I would recommend Firebase as a platform for other researchers looking to develop scalable online applications for data collection.

I used a combination of pseudonymisation, database security rules and Firebase authentication provided the security for Mindgames. However, webapp security is difficult to evaluate because it is impossible to know whether one's measures are insufficient until it is too late.

The platform met my requirements for anonymous login and does not require participants to signup, using their ProlificID to create an email-password pair and Firebase's authentication token UID to pseudonymise their data. However, this design feature may also be a security weakness. If a third party were able to guess an existing participant's ProlificID, then this would act as the email-password pair and they would be able to 'log in' as another participant: though the worst they could do would be to submit false experimental data. All sensitive data, such as cognitive data or questionnaire responses are write only. Furthermore, all data are pseudonymised, thus ensuring a second layer of privacy in the event of a data breach. Overall, I am confident of the level of data security Mindgames provides.

4.6 Chapter Summary

In this chapter I set out the design requirements of Mindgames, provided examples of how I addressed those requirements and briefly evaluated my implementation. As discussed in Chapter 3, I sought to develop my own platform for hosting web-based cognitive studies so

that I'd have full creative and technical control. This would enable me to run more complex, longitudinal, studies and to create higher quality gamification.

I had ten design requirements: sign-up free participation, support for longitudinal studies, game suitability, scalability, security, accurate response timing and accurate stimulus presentation timing, minimisation of cross-platform differences, extensibility and the provision of study management tools. These design requirements evolved over time as the prototype platform was used to run five studies. After each study I modified the platform to improve its functioning and correct bugs. The resulting platform is a single page webapp written in Javascript and HTML 5. To aid extensibility, I adopted an object-oriented design approach throughout. Mindgames was designed specifically to interface with Prolific.

I made extensive use of two libraries: Firebase and PixiJS. Firebase is a Google service platform + JavaScript library, which provides a JSON-like database alongside cloud-based web-hosting and advanced security features, it helped me deliver a scalable, secure application and made it easy to implement support for longitudinal studies and sign-up free participation. PixiJS is a 2D rendering engine which provides extensive support for drawing complex graphics on a HTML5 Canvas. PixiJS allowed me to deliver a consistent user experience across a range of browsers and made it easy to create game-like features such as animation and graphics.

Overall, I achieved my goal of creating an online platform for hosting cognitive tasks. I met my design requirements either by employing existing technology or developing my own. I took advantage of opportunities external to my thesis project to test and improve my platform, primarily in the areas of security, timing and ease of use for non-technical users. Most importantly, this platform gave me the ability to run complex longitudinal studies with hundreds of participants at once, and control over every aspect of the task. In Chapters 5 and 6, I use Mindgames to conduct two online studies through Prolific, with the goal of examining the effect of gamification on amount of engagement.

Chapter 5: The effects of points and theme on attrition from a web-based longitudinal cognitive testing study (Experiment 2)

This chapter is based on my publication in JMIR: [210]

5.1 Chapter Aims

In Experiment 1 (Chapter 3) I showed that the gamification of a cognitive task, specifically the addition of points, could improve quality of engagement. In Chapter 4 I described the development of Mindgames: a platform designed to host longitudinal studies of gamified cognitive testing. In this chapter I describe Experiment 2, run on that platform, which investigated whether the positive findings of Experiment 1 would transfer to amount of engagement. Specifically, could gamification reduce attrition from a multi-day cognitive testing study? I had four aims:

1. Investigate the effects of individual game elements of amount of engagement, measured using attrition.
2. Investigate the effects of individual game elements on quality of engagement
3. Investigate the effects of individual game elements on the primary outcome measure of the stop-signal task (Stop Signal Reaction Time)
4. Pilot two potential behavioural measures of engagement

5.2 Introduction

As discussed in Chapter 1, online studies (and particularly longitudinal studies) must compete against the wealth of entertainment and distraction available on the Internet to attract and retain their participants. Many authors have reported difficulties sustaining participant numbers for the duration of their online studies [10,11]. Reviews of adherence to intervention trials have documented levels of attrition as high as 50% [12,13], considerably higher than in laboratory studies where dropout rates are around 13% [211]. Participant attrition is a measure of amount of engagement, and is important because high dropout rates result in smaller than intended sample sizes, incomplete datasets, wasted participant compensation, and potentially biased results [15–17].

Two recent systematic reviews have looked at the effect of gamification on amount of engagement with ‘online programs’ (mostly e-learning) [43] and web-based mental health interventions [42]. Drawing on the data from 15 studies comparing engagement with gamified programs to non-gamified programs, Looyestyn and colleagues found medium to large effects of gamification on measures of amount of engagement such as time-spent using the program,

number of website visits and volume of contributions [43]. In contrast, Brown and colleagues assessed the impact of gamification on adherence to 61 online mental-health interventions and found that not only was gamification applied superficially (most studies used only one game element), there was also little evidence for its efficacy [42]. These conflicting findings could be the result of the reviews' different scopes, the lack of studies in Brown's review which specifically assessed the impact of gamification on adherence, or the primarily superficial gamification found to have been applied in the reviewed mental-health interventions.

Experiment 1 showed that the points variant was rated the highest of the three gamified task variants (non-game, points, theme) on a questionnaire of quality of engagement, while also not negatively affecting participant performance on the test. I found that the narratively themed task was less liked and negatively affected participant performance. I saw ceiling effects on participant accuracy in all three task variants, probably due to the ease of the response inhibition task.

In this study, I again used three variants of a response inhibition task but switched to using the stop-signal task (SST) to increase task difficulty and address ceiling effects. I used the same game design elements (non-game, points, and theme) and assessed their effect on attrition using a longitudinal design whereby participants signed-up to four compulsory test sessions over four consecutive days before entering a six-day voluntary period where they could continue to take part once per day if they desired. Participants were told they would receive £4 for completing all compulsory sessions and an additional 50p for each optional session they completed.

5.2.1 Hypotheses

I hypothesised that non-game variant would have the highest attrition rate, losing participants quickly after the compulsory sessions were complete. The non-game variant contained no game elements; its only obvious motivating factor was the reward of 50p per session. On the premise that points might suffer diminishing returns when used in isolation [154,212]; I expected the points variant to initially maintain high numbers, before falling rapidly around day 6-7. I hypothesised the theme variant would lose participants steadily at first before stabilising to a low attrition rate, eventually retaining a higher number of participants than either the non-game or points variants. I expected the map-screen would encourage participants to see the study through to completion by providing a sense of progression [123].

5.3 Methods

5.3.1 Design and Overview

I used a between-subjects, repeated measures experimental design that took place online over four to ten days. The independent variable was SST variant (non-game, points, theme). The dependent variables of interest were participant attrition, scores on a questionnaire of quality of engagement, two pilot behavioural measures of engagement and Stop-Signal Reaction Times (SSRTs). I pre-registered the study on the Open Science Framework (osf.io/58jur).

5.3.2 Participants and Procedure

Participants were recruited from the user base of Prolific, through which I handled the checking of inclusion criteria, displaying of study information and participant reimbursement. I required participants to be older than 18 and to have English as a first language, but had no further criteria. Once registered, participants were directed to Mindgames where they received a unique link which they used to access the study thereafter. They were then randomly assigned to a single task variant for the duration of the study and completed an online consent form before testing commenced.

Participants were required to complete one ten-minute session per day for the first four days of the study to receive £4 as compensation for their time. If participants dropped out of the study before completing four sessions and did not contact me with a reason (technical difficulties, etc.) then they did not receive any compensation. This was made clear on the information sheet which participants read before they signed up to the study, and on the study website itself. For the first four sessions participants were sent daily reminders via the Prolific messaging system. On the fourth day participants were informed there would be no more reminders, and that they were free to either drop out, or continue to take part in the study each day thereafter for up to six days, with each additional session earning them 50p, for a total of between £4 and £7.

The appropriate compensation for the optional sessions was determined by way of a pilot study using the non-game variant only. I randomly allocated participants to one of three levels of compensation: 50p, £1 or £2 per optional session completed (the base compensation was still £4) and found that the average number of sessions completed per participant was 7.1, 8.4 and 9.4 respectively. Given that I anticipated the non-game variant to be the least motivating of the three variants, that I wanted to avoid ceiling effects, and that I wanted to minimise the motivational influence of the compensation, I opted for a reward of 50p per optional session.

Ethics approval was obtained from the Faculty of Science Research Ethics Committee at the University of Bristol (40361) and the study was conducted according to the revised Declaration of Helsinki [171].

5.3.3 Materials

The Mindgames Platform

Aside from participant recruitment, daily reminders and reimbursement, all other elements of the study were delivered on Mindgames (Chapter 4). The site opened to a main-menu screen from which the participant could view the number of sessions they had completed and the amount of money they'd earned so far (Figure 5.1). Participants had access to a 'history' screen which allowed them to view their previous progress and monitor their results over time.

Clicking the start button displayed a series of instruction screens followed by the SST task and a short questionnaire. The session ended on the history screen, and the main menu's 'start' button became inactive until midnight that night. Each session took 10 minutes to complete. On the first day of taking part, participants also completed a short demographic questionnaire which collected data on age, sex, ethnicity, level of education and the number of hours spent playing video games each week.

Figure 5.1 Menu screens of the three task variants. (A) non-game variant, (B) points variant, (C) theme variant

Stop-Signal Task: non-game variant

The SST measures response inhibition [213,214] a key feature of executive control [215]. It tests the participant's 'action restraint' by presenting a series of stimuli to which the participant must respond as quickly as possible while occasionally being required to withhold a response [216]. These 'stop trials' are indicated by a visual warning presented a brief delay after stimulus presentation. The primary outcome measure of the SST is the SSRT which is the number of milliseconds of warning a participant needs for them to be able to have a 50% chance of inhibiting their planned response [215].

In this study I decided to use the SST as opposed to the Go/No-Go Task (GNG) used in Experiment 1. This was because many participants performed at ceiling in the GNG, which limited my ability to detect differences between the task variants. The SST is more challenging than the GNG because it dynamically adjusts the task's difficulty to match the inhibitory control of the user, therefore reducing the likelihood of a participant performing at ceiling.

I based my SST on the widely used CANTAB SST [217,218] albeit with a visual rather than auditory stop-signal and some graphical upgrades to make the task more suitable for online use. Each trial began with a fixation cross that was displayed in the middle of the screen, with two coloured zones on the left and right of the fixation cross (Figure 5.3A). 500 ms later a coloured circle appeared over the fixation cross and participants had to respond as rapidly as possible using their keyboard's left and right arrow keys to move the coloured circle to the matching side of the screen (right arrow for blue and left for yellow). On 25% of trials, white brackets appeared around the circle after it was shown (Figure 5.2): when this occurred the subject had to withhold their response and wait until the next trial began (each trial was displayed for 900 ms). If the participant responded before the stop-signal was displayed, then the trial was recorded as failed, but white brackets were not displayed. Between each trial there was a random inter-trial interval of between 500-1000 ms.

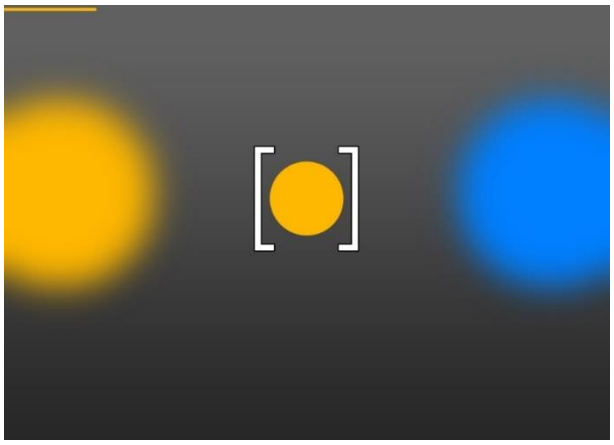


Figure 5.2 Screenshot of a stop-trial in the non-game variant of the SST. The white brackets around the stimulus indicate the participant should withhold their response.

The delay between the circle onset and the bracket onset is called the Stop-Signal Delay (SSD), and was varied according to a four-staircase tracking algorithm, designed to sample across the SSD/Inhibition-Probability space (Appendix K - pg154) [219,220]. The task consisted of five blocks of 48 trials each, with a 10 second break between each block. If the participant minimised the browser window or changed tabs then the task was paused. If the browser window lost focus but was still visible (on a 2nd monitor for example) then the task was not paused.

In the non-game variant, the participant's history was presented as a list of previous sessions, with median RTs and estimated SSRTs (Figure 5.3B). Hovering over a column displayed a brief explanation of the variable (e.g., "The Reaction Time column shows the average time, in milliseconds, which it took you to respond each session")

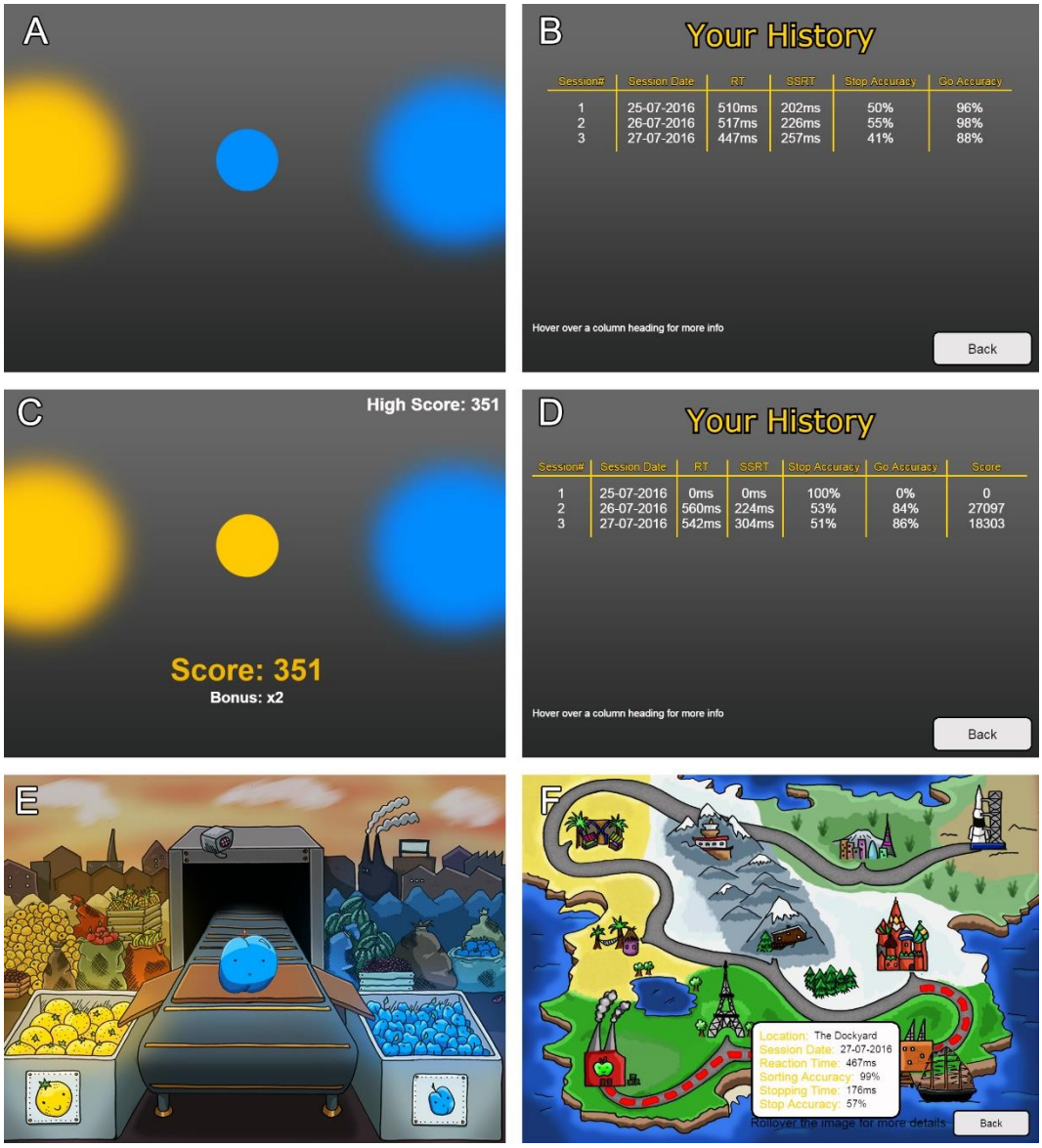


Figure 5.3 In-task screenshots of the SST variants and the associated history screens. (A/B) non-game variant, (C/D) points variant, (E/F) theme variant

Stop-Signal Task: points variant

The points variant was similar to the non-game variant but with the addition of a scoring system and the task being framed as a game (for full details, see Appendix J - pg146). In the task, the participant's points score was displayed at the bottom of the screen throughout (Figure 5.3C). The scoring system was very similar to that used in Experiment 1: on each successful non-stop-trial the participant earned points equal to $Bonus \times 0.2 \times (800 - \mathrm{RT})$, and the number of points gained was displayed briefly in the inter-trial interval. This Bonus

was a multiplier (x2, x3, x4...), which increased by 1 every 3 trials but decreased by 3 when the participant failed a stop trial. The bonus was not lost on stop trials to which the participant responded before the stop-signal was displayed (to all appearances, the trial was not a stop trial). On a successful inhibition to a stop-signal the bonus was not lost, but no points were awarded (as there was no RT on which to base the score for that trial). Scores were maintained over blocks, but not over sessions. The scoring system was outlined to the participants in the instructions for the task.

The participant's history was presented as a list of median RTs, SSRTs and scores from each session (Figure 5.3D). The participant's highest score was saved as a highscore and was displayed in the top right-hand corner throughout every testing session.

Stop-Signal Task: theme variant

The theme variant was similar to the non-game variant but with the addition of a graphical theme and a sense of progression (for full details, see Appendix J - pg146). The task was framed as a game and featured themed graphics and stimuli, with the yellow and blue stimuli replaced by images of blue and yellow objects (Figure 5.3E). The instructions provided a light narrative frame for the task, explaining that the player needed to sort the blue and yellow objects into their respective piles, but to avoid sorting any objects that were detected as 'faulty'. The task was presented on a series of different graphical backgrounds ('locations') which the player progressed through over the sessions of the study. Each location had several shared elements: a conveyor belt on which objects appeared and two bins to the left and right into which these objects were sorted.

The participant's history was presented as a map of the locations they'd visited (Figure 5.3F), and previous sessions' summary data was displayed when the user hovered over the corresponding icon. Each level on the map had a unique name and thematic instruction text, with the intention of creating an overarching goal, perceptual curiosity and fostering a sense of participant progression: all with the intention of increasing engagement [76,123,221].

Assessment of Quality of Engagement

After completing the task participants were presented with a brief questionnaire to assess their experience of the task. The same questions as in Experiment 1 (excepting item 11), were presented in a random order:

1. How enjoyable did you find the task?
2. How frustrating did you find the task?

3. Was it difficult to concentrate for the duration of the task?
4. How well do you think you performed on this task?
5. How mentally stimulating did you find this task to be?
6. How boring did you find the task?
7. How much effort did you put in throughout the task?
8. How repetitive was the task?
9. How willing would you be to take part in the study again?
10. How willing would you be to recommend the study to a friend?

The questionnaire was delivered after every session. Sessions 1,4,7 and 10 delivered the full ten-item questionnaire while the remaining sessions delivered a shorter five item questionnaire (questions 1,2,5,8 and 9). Items were answered using a continuous VAS, presented as a horizontal line 500 pixels long, with a label at either end and no subdivisions. Participants marked a point between these two labels using their mouse.

5.3.4 Dependent Variable Calculation

Attrition (Amount of Engagement)

Attrition was measured in two ways: Firstly, I calculated the mean number of sessions completed per participant (sessions which were started but not finished were excluded from this calculation). Secondly, I calculated the percentage of participants that completed at least one session, two sessions, etc.

Quality of Engagement

Participant quality of engagement with the task was measured by calculating a mean score from the 10-item assessment of quality of engagement. Questions 2,3,6 and 8 were reverse-scored. This measure was calculated for each participant's first and fourth sessions, and I also created a 'combined score' by taking the mean of the participant's scores from sessions 1 and 4.

Behavioural Measures of Engagement

I piloted two measures that I hoped might serve as proxies for engagement: I counted the number of times that participants hid the browser window or moved focus to another window while completing the SST, hypothesizing that unengaged participants would be more likely to briefly visit other websites while testing. I combined the counts of both these events into a single measure: loss-of-focus events. I then created an overall measure of loss-of-focus for each participant by calculating the mean number of loss-of-focus events from their first four sessions.

I also investigated coefficients of RT variation, which quantify RT intra-individual variability with respect to mean RT, as there is some evidence that changes in motivation can be reflected in RT variation [222,223]. Coefficients of variation were calculated for each session by dividing the standard deviation of non-stop trial RTs by the mean non-stop trial RT. I also created an overall measure of RT variation for each participant by calculating the mean coefficient of variation from their first four sessions.

Stop-Signal Reaction Times

I calculated SSRTs for each session separately, excluding sessions where the assumptions of the race model did not hold. The race model is a commonly used model of inhibitory control and aims to describe the relationship between stop and go processes (see [224] for overview). The race model is used to derive the SSRT and so if the assumptions underlying the race model are broken, then the resultant SSRTs are not good representations of the data [213,224]. To that end, I excluded sessions where the median non-stop-trial RT was longer than the median failed-stop trial RT, where SSDs were not positively correlated with their corresponding median failed-stop RTs, and where stop-trial accuracy was not negative correlated with SSD.

For the sessions which did meet the assumptions of the race model, SSRTs were calculated by modelling an inhibition function, and using it to estimate the SSD at which the participant's probability of inhibiting to a stop-signal was 50% [220], I then used this SSD to calculate the SSRT for that session [213,214]**.** I also created a combined measure of SSRT for each participant by taking the mean SSRT of their first four sessions.

The estimated SSRTs on the history screen were calculated automatically at the end of each session using the integration method [220].

5.3.5 Statistical Analysis

In all analyses, where appropriate, differences between groups were assessed with *post-hoc t*-tests. Where there was no evidence of a difference between group-means, I used Bayesian *t*-tests to assess the evidence for equality (3.3.5).

Attrition

Differences in attrition curves were assessed visually using the Kaplan Meier method to estimate survival functions, a Log-Rank test, and a one-way ANOVA of 'number of sessions completed'. In all cases the between-subjects factor was task variant (non-game, points, theme).

Quality of Engagement

Quality of engagement was assessed both visually, using bar-charts, and using a repeated-measures ANOVA of total score with session number as the time factor (Session 1, Session 4) and task variant (non-game, points, theme) as a between-subjects factor.

Behavioural Measures of Engagement

I assessed differences in coefficients of variation and loss of focus events between task variants using one-way ANOVAs with data combined across the first four sessions, and task variant (non-game, points, theme) as a between-subjects factor.

Stop-Signal Reaction Times

I assessed the effects of gamification on SSRT using boxplots and a one-way ANOVA of SSRT with a between-subjects factor of task variant (non-game, points, theme).

5.3.6 Sample Size Determination

At the time of study design, to the best of my knowledge, no other studies had investigated the impact of gamification on attrition from a cognitive testing regime, and therefore I had no previous effect size on which to base a sample size determination. Instead, I hypothesized attrition curves (Figure 5.4) for each variant and calculated the anticipated effect size (ϕ=.231) resulting from a Kaplan-Meier method/Log-rank test (i.e., a chi-square test) on those attrition curves. To detect this difference with alpha = 0.05 and 95% power a sample size of 290 was required. I set this to 291 to allow for equal group sizes.

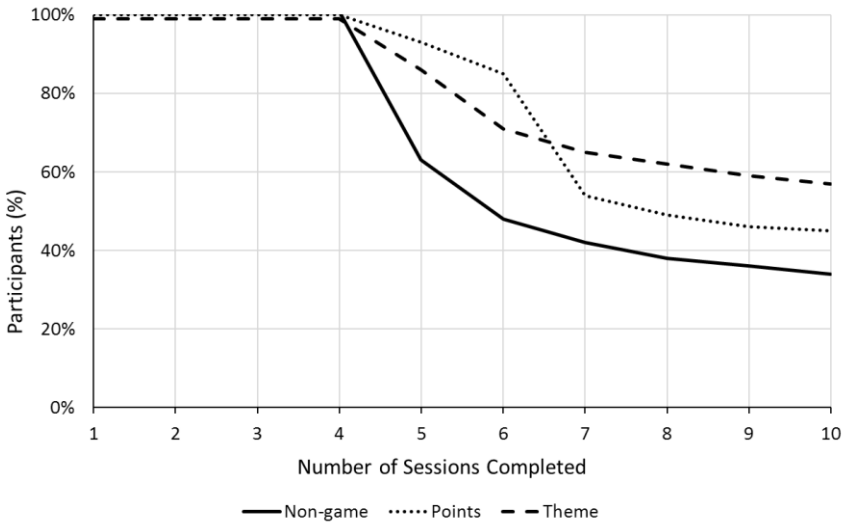


Figure 5.4 Percentage of participants plotted against the number of sessions I hypothesised they would complete, shown separately by task variant.

5.4 Results

The data that form the basis of these results are available on request from the University of Bristol Research Data Repository (10.5523/bris.2lh5oz3p93q7p2uyx4w2h70a0v)

5.4.1 Characteristics of Participants

Participants were recruited in two waves: one starting October 2016 and another starting in January 2017. In both waves the intended sample size was met within three days of the study being posted on Prolific. A total of 482 participants signed up to take part in the study, with 419 (86.9%) of those completing at least one session. A total of 265 (54.9%) participants completed four sessions over four consecutive days as was required by the study criteria (henceforth called *conforming participants*). I excluded five participants from the main analysis because their RTs or blue/yellow accuracy scores were more than four interquartile ranges away from the group median. I excluded data from sessions that were started but not completed, and I removed trials from the analysis where participants responded in less than 150ms.

Excluding outliers, 260 conforming participants took part: 91 in the non-game variant, 86 in the points variant and 83 in the theme. The number of hours spent playing video games was comparable between the groups, and participants typically had a high level of education (Table 5.1). The most common browser used to complete the experiment was Google Chrome (71%), with others including Firefox (19%), Netscape (5%), Safari (4%), Opera (0.5%) and Internet Explorer (0.5%).

Table 5.1 Conforming participant demographic information, shown separate by task variant.

	Non-game	**Points**	**Theme**
Mean age (SD)	36 (12)	35 (12)	34 (11)
% Male	47%	57%	51%
Mean video game hours per week (SD)	6 (12)	8 (16)	8 (14)
Median level of education	Bachelor's Degree	Bachelor's Degree	Bachelor's Degree
Mode ethnicity (percentage)	Caucasian (88%)	Caucasian (86%)	Caucasian (90%)

As the study was underway, it became apparent that 32 participants were unable to complete the required four sessions in four days, but instead managed to complete four sessions within five days. During the study, I intended to include these *loosely conforming* participants in the analysis, and so stopped recruitment once my intended sample size was achieved. For simplicity and adherence to the protocol, most of analyses below present data from the 260 conforming participants only. However, I present an attrition analysis for both conforming and

loosely-conforming participants. Full analyses including all participants can be found in Appendix L (pg155).

5.4.2 Conforming Participant Attrition

Figure 5.5 shows the attrition of conforming participants, while Table 5.2 shows the mean number of sessions completed per participant in each variant. A Log-Rank test showed no evidence of a difference between the distributions ($X^2_{2,260}$=2.460, p=.29, ϕ=.097) and a one-way ANOVA of the number of sessions completed also found no clear evidence of a difference between task variants ($F_{2,259}$=1.360, p=.26, *partial* η^2 =.010). Given the similarity between non-game and points in mean number of sessions completed, I used a Bayesian *t*-test to assess their equality and found substantial evidence that they were equal (Bayes Factor (BF)=.16), but there was no evidence of equality between the theme and the points variant (BF=.49) or the non-game variant (BF=.43).

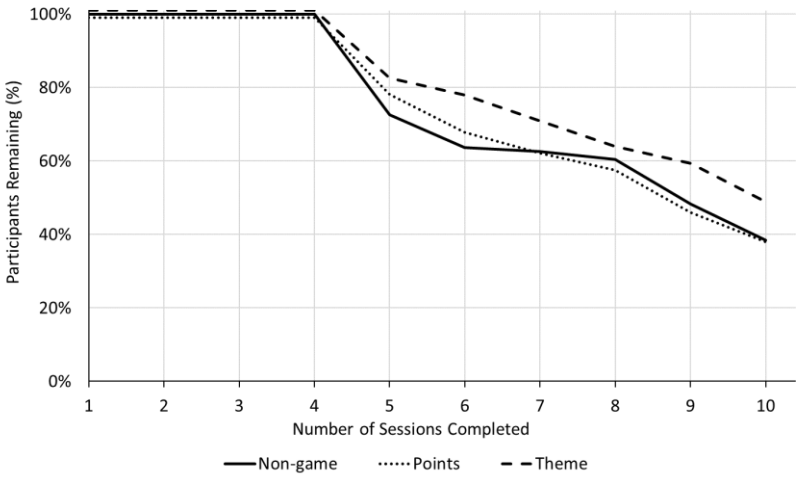


Figure 5.5 Percentage of conforming participants plotted against the number of sessions they completed, shown separately by task variant.

Table 5.2 Mean number of sessions completed per participant, shown separately by task variant. Conforming participants are those who completed their first four sessions within four days as required. 'All participants' includes all who signed up, regardless of their number of sessions completed.

	All participants (95% CI)	**Conforming participants (95% CI)**
Non-game	4.9 (4.4 to 5.5)	7.4 (6.8 to 8.0)
Points	5.1 (4.5 to 5.6)	7.5 (7.0 to 8.0)
Theme	5.3 (4.7 to 5.9)	8.0 (7.5 to 8.6)

5.4.3 Loosely-Conforming Participant Attrition

Table 5.3 and Figure 5.6 show the attrition of loosely conforming participants. This includes 260 conforming participants in addition to 32 participants who only managed to complete their final compulsory test session on the fifth day of the study, rather than the fourth. Given the study's 10-day duration, the maximum number of sessions that *all* participants had a chance to complete was 9.

Table 5.3 Mean number of sessions completed within 9 days, shown separately by task variant. Conforming participants are those who completed their first four sessions within four days as required. Loosely conforming participants includes conforming participants AND participants who completed their first four sessions within five days.

	All participants (95% CI)	**Conforming participants (95% CI)**	**Loosely conforming participants (95% CI)**
Non-game	4.9 (4.4 to 5.5)	7.0 (6.6 to 7.5)	6.9 (6.5 to 7.4)
Points	5.1 (4.5 to 5.6)	7.1 (6.7 to 7.6)	7.0 (6.6 to 7.4)
Theme	5.3 (4.7 to 5.9)	7.6 (7.1 to 8.0)	7.3 (6.8 to 7.7)

I used the Kaplan Meier method to calculate estimated survival times of loosely conforming participants. A Log-rank test showed no evidence of a difference between the distributions ($X^2_{2,292}$=1.082, p=.58) and a one-way ANOVA of the number of sessions completed also found no evidence of a difference between task variants ($F_{2,291}$=.544, p=.58, *partial* η^2=.004).

Bayesian *t*-tests found substantial evidence that the number of sessions completed in all variants was equal, with points being equal to non-game (BF=.16), points being equal to theme (BF=.23) and non-game being equal to theme (BF=.25).

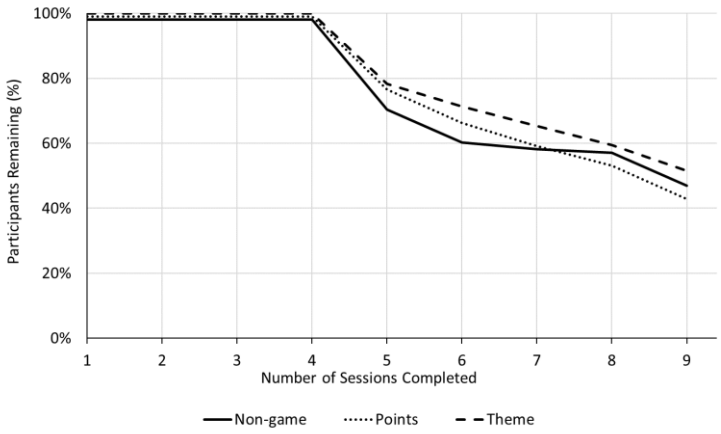


Figure 5.6 Percentage of loosely conforming participants plotted against the number of sessions they completed, shown separately by task variant.

5.4.4 Quality of Engagement

I used a repeated-measures ANOVA of mean score from the assessment of quality of engagement with session number (1,4) as the within-subjects factor, and task variant as the

between. I used only the two full-length questionnaires completed on the 1st and the 4th session, completed by all participants. I saw evidence for small effects of both task variant ($F_{2,259}$=3.805, p=.02, *partial* η^2=.028) and time ($F_{1,260}$=35.693, p<.001, *partial* η^2=.120), and weak evidence of an interaction ($F_{2,259}$ = 3.014, p=.05, *partial* η^2=.023). Ratings of all task variants decreased between the first (M=56, 95% CI 54 to 57) and fourth session (M = 51, 95% CI 49 to 53), but it appears the non-game and points variants were the main drivers of the interaction effect: dropping by 6% (95% CI 4% to 8%) between Session 1 and Session 4, whereas ratings of the theme task decreased by only 2% (95% CI -1% to 5%). Post-hoc *t*-tests on the combined scores showed no evidence for differences between non-game and points, nor non-game and theme (*ps*>.15), but did show points and theme to be different (mean difference=5.7%, 95% CI 1.6 to 9.7, t_{169}=2.749, p=.007, d=.42). Figure 5.7 shows the mean scores from each task variant at the two time points, and a combined score from the mean of both sessions. A breakdown of ratings by individual questions is presented in Appendix M.

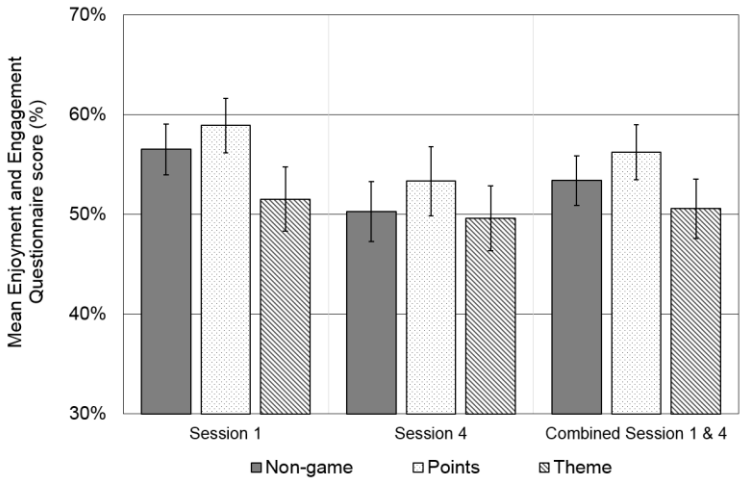


Figure 5.7 Overall scores from the assessment of quality of engagement. Mean responses of visual-analogue scale scores from questionnaires delivered on sessions 1 and 4, and the mean of scores from Sessions 1 and 4, shown separately by task variant and time point. Error bars represent 95% CIs

As an unplanned exploratory analysis, I was interested to see whether quality of engagement was associated with amount of engagement. I ran a Pearson's correlation and found weak evidence that the combined scores of the assessment of quality of engagement were positively associated with the number of sessions completed (r=.116, p=.062). I further explored the relationship to determine whether a participant's rating one day predicted their return to the study on the following day. I ran a logistic regression with "returned following day" as the binary dependent variable and the previous day's score on the questionnaire as the predictor variable. However, there was no evidence that assessment of quality of engagement scores

predicted return the following day, (β=0.008, SE=.005, Wald(1)=2.166, p=.14, Odds Ratio=1.001, 95% CI 0.997 to 1.019)

5.4.5 Behavioural Measures of Engagement

I analysed RT coefficient of variation and website loss of focus events from the four compulsory sessions combined (Table 5.4). A one-way ANOVA of coefficient of variation showed strong evidence for a medium effect of task variant ($F_{2,259}$ = 3.131, p=.045, *partial* η^2=.024) on participants' RT variability, with lower coefficients indicating there was less variability. Post-hoc *t*-tests showed strong evidence of a small difference between the points and theme variants (with theme being less variable) (mean difference = 1.5%, 95% CI 0.2 to 2.7, t_{169}=2.349, p=.02, d=.36), but no clear evidence for other differences were found (ps>.06).

Loss-of-focus events were rare in all task variants, with each participant switching away from the task less than once per session on average. Regardless, I assessed differences in loss-of-focus events between the three task variants using a one-way ANOVA but found no evidence for any differences ($F_{2,259}$=1.137, p=.32, *partial* η^2=.008).

Table 5.4 Mean behavioural measures of participant engagement from the first four sessions, shown separately by task variant.

	Coefficient of variation (95% CI)	**Loss-of-focus events (95% CI)**
Non-game	18.7% (17.9 to 19.6)	0.85 (0.50 to 1.19)
Points	19.0% (18.1 to 19.8)	0.82 (0.43 to 1.20)
Theme	17.5% (16.7 to 18.4)	1.21 (0.75 to 1.67)

5.4.6 Stop-Signal Reaction Times

I checked the data from each session against the assumptions of the race model. Of the 1050 sessions assessed, I excluded 161: 75 from the non-game variant, 37 from points and 49 from theme. 3 participants failed to meet the assumptions of the race model in all four compulsory sessions, resulting in their exclusion from this analysis. I then analysed each participant's mean SSRT, with boxplots shown in Figure 5.8.

A one-way ANOVA showed weak evidence for a small effect of task variant on SSRT ($F_{2,255}$=2.954, p=.05, *partial* η^2=.022) with post-hoc *t*-tests showing a difference between the theme variant (M = 289, SD = 67) and points variant (M=266, SD = 66) (mean difference = 23, 95% CI 5 to 42, t_{169}=2.386, p=.05, d=.35). There was no evidence for other differences (ps>.24). Bayesian *t*-tests showed no evidence of equality between the SSRTs of the non-game and theme variants (BF=.59), but found substantial evidence for equality between the non-game (M=274, SD=55) and the points variants (BF=.22).

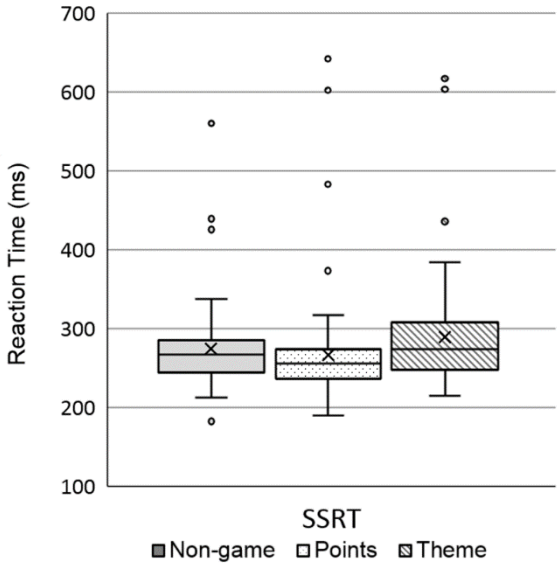


Figure 5.8 Boxplots of mean SSRT. Data combined per participant over the first four sessions and shown separately by task variant

For brevity, not all the analyses planned in the study protocol have been presented– for more detailed methods and analyses please see Appendix N (pg157).

5.5 Discussion

Contrary to my hypotheses, I found no clear evidence of an effect of task variant on participant attrition. This was further strengthened when I included data from loosely conforming participants which showed strong evidence that the mean number of sessions completed was equal in all task variants. To the best of my knowledge, this is the first empirical study examining the effects of gamification on participant attrition within a cognitive testing context, and my results raise doubts about the effectiveness of gamified tasks for increasing amount of engagement.

5.5.1 Quality of Engagement

Despite there being no difference in amount of engagement between the variants, I did find an effect of task variant on quality of engagement, with the points variant having the highest combined sessions mean, followed by the non-game variant and the theme. One possible explanation for these findings relates to self-determination theory (SDT) (Section 2.4.4). In the case of the gamified variants, the points variant would seem to address competency needs by providing constant feedback on their performance which reinforces the player's success [37], but I would not consider the theme variant or the non-game variant to adequately meet any of the three needs. Since the points variant was the only variant to address any of the psychological needs, this might explain why it was rated as providing the highest quality of engagement in both this study and in Experiment 1.

The theme variant was rated as providing the lowest quality of engagement, even compared against the non-game variant. One potential explanation for this finding is that the task was framed as a game and looked like a game, but offered no actual gameplay. Secondly, the map screen and changing graphical backgrounds may have hinted at player autonomy and exploration as is typical in other games, but ultimately the player experience was railroaded. These two factors may have undermined autonomy and violated participant expectations, resulting in a dissatisfying experience [134,225].

One additional factor to consider, in the light of SDT, is that paying participants in studies such as this may be counterproductive to measuring true engagement. There is evidence that providing extrinsic rewards for otherwise motivating tasks may undermine participant autonomy, therefore affecting the task's ability to meet our psychological needs [226,227]. In this study, it is not possible to determine whether intrinsic motivation to take part was affected by the incentive of 50p per additional session. This is further complicated by the potentially unrepresentative nature of a Prolific sample: all of whom have voluntarily signed up to take part in science experiments online, and can choose studies based on the amount of monetary compensation awarded in exchange for their data. Given these issues, an informative avenue for future research in this area would be to explore the effects of these same gamification mechanisms on attrition, but without providing financial incentives. Money can be a powerful motivator; for example, Khadjesari and colleagues [228] found that offering a £10 Amazon voucher to each participant in a longitudinal study resulted in a 9% increased response rate at 12-month follow-up. In this case, it may simply be that money was the most important factor for taking part, and that the similar attrition rates were driven by the identical incentives.

I also found only very weak evidence that quality of engagement was associated with amount of engagement. This highlights the different types of engagement that Perski and colleagues found in their review of the concept of engagement [150]. It is assumed that measures of quality of engagement and amount of engagement are triangulating the same concept, but evidence in support of this claim is scarce. Research from the video game literature has found that game enjoyment does not relate strongly to game usage, and that game usage can be driven by many other factors including boredom, loneliness and need for escapism [149,229]. This evidence, combined with the findings of this study, indicates that further research is needed to understand the relationship between quality and amount of engagement.

5.5.2 Cognitive Data

The two pilot behavioural measures of engagement: RT variation (coefficients of variation) and loss-of-focus events were difficult to interpret. There was no evidence that losses of focus differed between the task variants, and this is likely because such events were rare (less than one loss of focus per session on average). However, this is a positive finding as it shows that participants are willing to properly engage with online cognitive tasks, concentrating for the duration. With respect to coefficients of variation, the pattern of results is directly in contrast with the quality of engagement: the points variant had the most variable response times but the highest rating, while the theme variant had the lowest variability and the lowest rating. This is either contrary evidence to the idea that RT variability is related to motivation [222,223], or signals that the questionnaire ratings are not good measures of motivation. Regardless, further research is necessary to understand whether these measures reflect engagement.

When assessing cognitive data, I found evidence that SSRTs were equivalent between the points variant and the non-game variant. Although the points variant introduced additional elements to the task which may have increased cognitive load, it is possible that the salient feedback and motivational effect of points served to counteract this: boosting participant performance as has been found in a number of previous studies [32,212,230,231].

5.5.3 Limitations

I acknowledge several limitations to this study: firstly, I did not technically achieve my intended sample size of 291. However, the attrition analysis of loosely conforming participants strengthens the finding of the main analysis: that there was no effect of gamification on attrition. Nevertheless, I accept that a balanced group analysis would be preferable. Secondly, I acknowledge that the design of the study was not suitable to validate the gamified variants as measures of response inhibition, as that would require a within-subjects design in order to test predictive validity [152,198]. Furthermore, a within-subjects design would reduce noise in the measurement of engagement by allowing participants to serve as their own control. Thirdly, it may be that my measure of amount of engagement was insensitive. I required participants to return to a study website and complete a ten-minute task in order to register one point on my measure of amount of engagement. It may be that the effect of gamification is not large enough to encourage this not-insignificant effort. Fourthly, as mentioned previously, there are issues relating to motivation and incentives, as in reality participants completing cognitive assessments will be requested to complete the study over a fixed period for a fixed fee, and not with the option to continue for additional recompense.

Fifthly, I acknowledge that the sample, recruited from Prolific, and with high levels of education, may not be representative of the wider population. Sixthly, this study investigated attrition over a period of six days, so results may not generalise to longer timecourses, such as attrition over weeks or months. Finally, the game elements I implemented were relatively superficial, and certainly wouldn't constitute a full game. Indeed, neither of the games were likely enjoyable enough that a participant would consider doing them for their own sake. Though this was necessary to try to reduce the impact of gamification on the cognitive data, it may have diminished any potential effects of gamification.

5.6 Chapter Summary

This chapter described Experiment 2: an empirical study into the effect of gamification on attrition from a longitudinal cognitive testing study. The study was hosted on Mindgames, and participants signed up to complete four consecutive days of testing, followed by six optional days. Each day, participants completed a 10-minute SST followed by a brief questionnaire on quality of engagement. Participants were randomly allocated to one of three task variants (non-game, points, theme). I used participant attrition over the six optional days as a measure of amount of engagement and hypothesised that participants would drop out of the study as they decided that completing the task was no longer worth 50p.

Contrary to my hypotheses, I found no evidence for an effect of gamification on amount of engagement. The theme variant had negative effects on cognitive data, showed no evidence of reducing attrition, was rated as the least enjoyable and was the task switched away from most often. This suggests that themed gamified tasks, at least those that use graphics alone, are non-optimal for use in cognitive testing studies. In contrast, and replicating the finding of Experiment 1, the points variant was found to provide the highest quality of engagement. I found SSRTs from the points and non-game variants to be equal, showing that points can be an effective way of improving quality of engagement with a cognitive task while still collecting valid data.

My findings show that there is still further work to be done untangling the relationship (or lack of one) between quality and amount of engagement; and further research should examine the role that specific game elements play. I acknowledge several limitations to Experiment 2, indicating scope for improvement with respect to experimental design. In the next chapter I describe Experiment 3, which takes an alternative approach to measuring amount of engagement and seeks to address limitations one to four.

Chapter 6: The effects of points, theme and financial incentive on amount of engagement (Experiment 3)

6.1 Chapter Aims

In Chapter 5 I looked at the effect of gamification on attrition from a longitudinal cognitive testing study and found that despite some positive effects on quality of engagement, there was no difference in dropout rates between the three task variants. There were, however, several methodological problems which might have diminished any potential effect. In this chapter I describe Experiment 3, which used Mindgames to investigate the effect of gamification on amount of engagement over three variable-length testing sessions, as opposed to attrition over a several-day period. I had four aims, namely to:

1. Investigate the effects of individual game elements of amount of engagement, measured by time spent testing.
2. Investigate the relationship between amount and quality of engagement
3. Investigate the effect of financial incentive on engagement.
4. Investigate the effects of individual game elements on the primary cognitive outcome measure of the stop-signal task (Stop Signal Reaction Time), using a within-subjects design

6.2 Introduction

In Chapters 2 and 5 of this thesis I have highlighted the lack of evidence that gamification can increase amount of engagement with cognitive tests. To the best of my knowledge, only a handful of studies have explored the issue: but these initial findings do suggest that gamification can increase the number of trials voluntarily completed by participants [50,52,232].

This chapter describes a study into the effects of game elements and financial incentive on the engagement with a cognitive test. I used a mixed design where participants completed three test sessions over a five-day period. Each session used a different variant of the stop-signal task (SST) (non-game, points and theme) in a counterbalanced order. The task was followed by two short questionnaires on the participant's quality of engagement. This experiment was comparable to Experiment 2 in many respects but had five important differences.

Firstly, I moved to a within-subjects design, I hoped this would provide better measures of engagement because participants would be able to compare their experience across the three tasks. It also allowed me to account for individual differences in my analysis, giving me more statistical power to detect effects.

Secondly, I switched from attrition as a measure of amount of engagement to *ad-libitum* test time [233]. Each session was of variable duration, with the SST being paused every two

minutes to ask participants if they wished to continue testing. I expected *ad-libitum* test-time to be a more sensitive measure of engagement than *number of sessions completed*, as the effort required to continue engaging was much less. i.e., continuation only required a button click and two more minutes of testing; as opposed to waiting a day, returning to the study website and completing a whole ten-minute test session.

Thirdly, I abandoned the somewhat unrefined assessment of quality of engagement used in Experiments 1 and 2, and I instead used the newly developed Digital Behaviour Change Intervention (DBCI) Engagement Scale. Though still in the process of being validated, I considered its design to be more grounded in theory than my own questionnaire. The DBCI Engagement Scale measures participants' quality of engagement with digital behaviour change interventions. DBCIs often make use of gamification to increase user engagement, hence I considered the DBCI Engagement Scale appropriate for use in this study.

Fourthly, to understand whether self-determination theory (SDT) (Section 2.4.4) could explain points' and theme's effects on quality of engagement, every session included a delivery of the Intrinsic Motivation Inventory (IMI). SDT posits that an activity is more appealing if it facilitates high levels of competency and autonomy in the individual [133,135]. I used a shortened version of the IMI to measure two factors: the participant's *perceived autonomy* and *perceived competence* during the task.

Fifthly, an important limitation of the Experiment 2 was that paying participants to complete additional sessions may have unintentionally influenced their engagement. There is evidence that providing extrinsic rewards for otherwise motivating tasks may undermine participant autonomy and affect the task's ability to meet psychological needs [226,227]. It may also be that the motivational effect of financial incentive was so large that it overpowered any possible effect of gamification. Gamification has been used in online studies (where financial incentives are common) and in DBCIs (where they are not). As such, it was important to understand how gamification affects engagement when used in situations both with and without financial incentive.

To include financial incentive as a factor in the study design: participants were recruited in two cohorts, each using different reimbursement schemes. One cohort received payment of £1.25 for completing each session regardless of time spent, while the other was incentivised to complete each block in exchange for a linearly decreasing financial reward. Under both reimbursement schemes, each session included one compulsory, two-minute, block of testing for which participants were paid £0.75. Following each block participants were asked if they

wished to continue and complete another block. I expected participants to drop out of the study when they decided the incentive (or lack of it) was not worth another two minutes of their time. All participants were additionally paid 50p for completing the compulsory post-task questionnaires each session.

6.2.1 Hypotheses

I hypothesised that the amount of engagement would be highest in the two gamified variants, and lower in the flat-rate reimbursement scheme. I also hypothesised that the points variant would be rated as the most engaging of the three, and that the DBCI engagement scale would be positively associated with amount of engagement. Finally, I hypothesised that the points variant would provide the most *perceived competency* on the IMI and the theme variant would provide the most *perceived autonomy.*

6.3 Methods

6.3.1 Design and Overview

I used a three-session 2 × 3 crossover experimental design, with a between-subjects factor of reimbursement scheme (flat-rate, pay-per-block) and a within-subjects factor of SST variant (non-game, points, theme). The dependent variables of interest were amount of engagement (defined as minutes spent on the task), quality of engagement (measured by the DBCI Engagement scale), intrinsic motivation (measured by the IMI), coefficients of RT variation and SSRTs. I preregistered the study on the Open Science Framework (osf.io/fytxj).

6.3.2 Participants and Procedure

Participants were recruited from the user base of Prolific, through which I managed the process of recruiting participants, checking inclusion criteria, displaying study information and participant reimbursement. I required participants to be older than 18 years, have English as a first language and have an approval rate above 90% on Prolific (as is recommended for longitudinal studies). I also required that participants had not taken part in Experiment 2.

Due to technical limitations on both Mindgames and Prolific's part, the two cohorts were run at separate timepoints. Which cohort a participant joined depended on the day which they signed up to the study. To minimise potential recruitment-pool differences given the non-random cohort allocation, I launched both cohorts on the same day of the week, at the same time, one week apart. The pay-per-block cohort was launched first. Once registered, participants were directed to *Mindgames* via a unique link, they were then randomly allocated to a counterbalanced task-variant order and an initial task variant. They were required to complete a consent form before they entered the experiment proper.

Participants were required to complete all three sessions within a five-day period to be paid. Participants could only complete one session per day and were invited to each session separately. If they dropped out of the study before completing three sessions and did not contact me with a reason (technical difficulties, etc), they did not receive any compensation. This was made clear on the information sheet, which participants read before they signed up to the study, and on the study website itself.

Each session rewarded a minimum of £1.25 + any bonus payments earned, up to a maximum of £4.01. Participants in the flat-rate reimbursement scheme were offered no additional financial incentive for completing additional blocks, while participants in the pay-per-block reimbursement scheme were offered a linearly decreasing incentive (66p, 57p, 48p, 39p, 30p, 21p, 12p and 3p) to complete each additional block (based on [234]) (See Table 6.1 for clarification). The total compensation a participant could receive was £3.75-£12.03. Ethics approval was obtained from the Faculty of Science Research Ethics Committee at the University of Bristol (60461), and the study was conducted according to the revised Declaration of Helsinki [171].

Table 6.1 The relationship between time spent testing, number of blocks completed and participant rewards for each session, in both cohorts. Participants could complete any number of blocks in any of the three sessions they completed. Completing one block + the questionnaires was compulsory in order for the session to be considered complete.

Time spent testing in a single session	**Blocks completed**	**Block rewards in the pay-per-block cohort (cumulative earnings)**	**Block rewards in the flat-rate cohort (cumulative earnings)**
2 minutes	1 (compulsory)	£0.75 (£0.75)	£0.75 (£0.75)
4 minutes	2	£0.66 (£1.41)	£0 (£0.75)
6 minutes	3	£0.57 (£1.98)	£0 (£0.75)
8 minutes	4	£0.48 (£2.46)	£0 (£0.75)
10 minutes	5	£0.39 (£2.85)	£0 (£0.75)
12 minutes	6	£0.30 (£3.15)	£0 (£0.75)
14 minutes	7	£0.21 (£3.36)	£0 (£0.75)
16 minutes	8	£0.12 (£3.48)	£0 (£0.75)
18 minutes	9	£0.03 (£3.51)	£0 (£0.75)
20 minutes	Questionnaires (compulsory)	£0.50 (£1.25 - £4.01)	£0.50 (£1.25)

6.3.3 Materials

The Mindgames Platform

Aside from participant recruitment, daily reminders and reimbursement, all other elements of the study were hosted on Mindgames (Chapter 4). The site opened to a menu from which the participant could begin the task or review the instructions. Clicking the start button displayed a series of instruction screens, followed by the variable-length SST and both questionnaires. On

the first day, participants completed a demographic questionnaire, which collected data on age, sex, ethnicity, level of education, and the number of hours spent playing video games each week. Each session took 4-20 minutes to complete, and once completed the main menu's *start* button became inactive until midnight.

Stop-Signal Task: Nongame Variant

I reused the SST from Experiment 2 (5.3.2) with some minor modifications. Firstly, I shortened the length of each trial to squeeze more trials into a 2-minute block, to improve the accuracy of SSRT estimates for participants who only completed one block of testing before dropping out. I reduced the response period of each trial from 900ms to 800ms, and I reduced the duration of the inter-trial interval from between 500-1000ms to between 200-500ms. I increased the number of trials in a block from 48 to 64.

I adjusted the stop-signal delay (SSD) staircase algorithm to be more conventional compared to that used in Experiment 2 (Appendix K - pg154). In the new design, all four staircases converged to a 50% failed inhibition rate, but each started from a different initial SSD (50ms, 125ms, 225ms, 300ms). On a step up or step down the staircase was adjusted by +/-50 ms respectively, the step size changed to +/-25ms after 3 reversals and to +/-12ms after 5 reversals. Staircase SSD values were maintained across blocks, meaning SSRT estimates became more accurate the longer the participant tested for.

At the end of a block a choice screen was displayed (Figure 6.1). The participant was told: "You are free to end today's session now if you wish. Alternatively, you may complete another two-minute round of testing [and earn an additional £X]. Would you like to continue?", where X was determined by the number of blocks completed, and the text in square brackets was only shown to participants in the pay-per-block reimbursement scheme. If the participant continued, they were presented with a short break screen for 10 seconds, after which the task continued. If they quit, they were directed to the questionnaires. Each participant could complete up to 9 blocks per session, with each block taking 2 minutes.

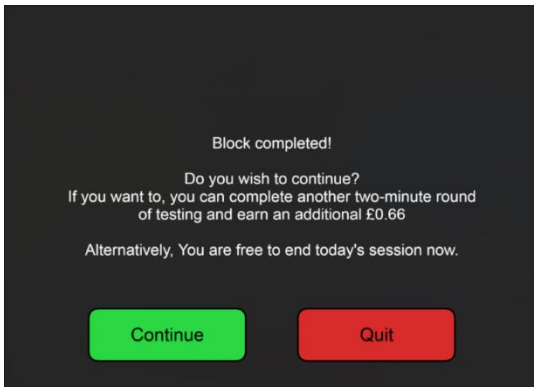


Figure 6.1 Screenshot of the choice screen from the pay-per-block reimbursement scheme

Stop-Signal Task: Points Variant

The points variant was similar to its counterpart in Experiment 2 (for full details, see Appendix J - pg146). However, the participant's score was not maintained over blocks, meaning they could compete against themselves on each block to beat their own high score. Given my decision to maintain staircase SSDs over blocks, I was concerned that it would become increasingly difficult for a participant to beat their own high score as the task went on (as the SST converged on their 50% failed inhibition rate). To counteract this, I included a *LevelAdjustment* variable in the score calculation, which served to increase the points earned by the participant by 5% for each block they had completed. The result was that on each successful non-stop-trial the participant earned points equal to $LevelAdjustment \times Bonus \times 0.2 \times (800 - reaction\ time)$.

At the end of each block the choice screen was presented (Figure 6.1) (identical to the choice screen in the non-game task variant). If the participant chose to continue taking part they were presented with a break screen for ten seconds, which displayed their current high-score and their score from the previous round. This break screen challenged the participant to beat their current high score in the next round (Figure 6.2).

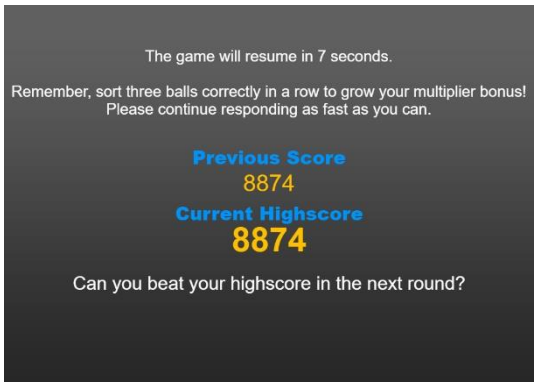


Figure 6.2 Screenshot of the post-continuation choice screen from the points variant of the SST, challenging the participant to beat their current highscore in the next round

Stop-Signal Task: Theme Variant

The theme variant was also similar to its counterpart in Experiment 2 (for full details, see Appendix J - pg146). If the participant chose to continue testing beyond the initial compulsory block (Figure 6.1)., then they were taken to a new choice screen. On this screen, they were given a choice of the location they would like to 'travel to next' (Figure 6.3), in effect giving them control over the graphical theme of their next block. My aim was to create a sense of autonomy: giving the participant some freedom in the way they completed the session [133]. Once they made their choice, they were presented with a self-paced break screen, which showed a paragraph of 'flavour text' about their selected destination.

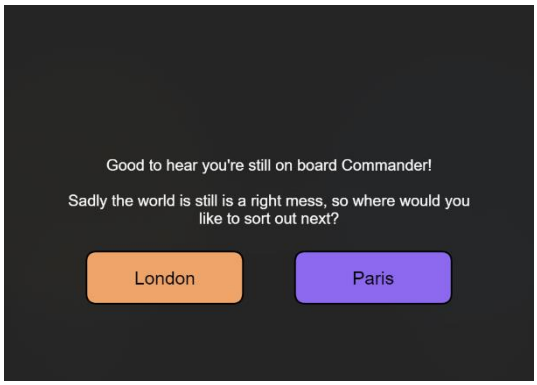


Figure 6.3 Screenshot of the post-continuation choice screen from the theme variant of the SST, asking the participant which location they'd like to 'sort out' next.

DBCI Engagement Scale

The DBCI Engagement Scale is designed to collect information on participants' engagement with DBCIs. DBCIs often make use of gamification to increase user engagement, hence I considered the scale appropriate for use in this study.

A shortened version of the DBCI Engagement Scale was presented after each testing session. The questionnaire consisted of 8 items presented in a random order and measured on a 7-point Likert scale with end points and middle anchored: not at all, moderately. extremely. The questions were: How strongly did you experience the following? (1) Interest, (2), Intrigue, (3), Focus, (4), Inattention, (5), Distraction, (6), Enjoyment, (7) Annoyance, (8) Pleasure.

Intrinsic Motivation Inventory

The IMI is a multidimensional questionnaire intended to measure participants' subjective experience of various activities [138,235]. The IMI contains many possible questions and subscales from which researchers select a subset to measure their factors of interest. I selected three questions designed to measure 'perceived autonomy' and three questions designed to measure 'perceived competence'.

The IMI was presented after each testing session, following the DBCI Engagement Scale. The questionnaire consisted of 6 items measured on a 7-point Likert scale with end points and middle anchored: not at all; moderately; extremely. The following questions were presented in a random order: How much is the following statement true? (1) I had some choice in how I approached this task, (2) I took part for as long as I did because I wanted to, (3) I felt like it was not my own choice to take part for as long as I did, (4), I performed well on this task, (5) It was clear how well I was performing on the task, (6) I was not skilled at this task.

6.3.4 Dependent Variable Calculation

Amount of Engagement

Amount of Engagement was measured in minutes: calculated as 2 × the number of blocks completed (each block was 2 minutes long). It was measured on a per session basis.

Quality of Engagement

Quality of engagement was measured using the DBCI Engagement Scale on a per session basis. I created three measures from this questionnaire: Interest, Attention and Enjoyment. Each measure was derived by calculating the mean score of its component items for each session. The Interest subscale consisted of items 1 and 2. The Attention subscale consisted of items 3, 4 (reversed-scored) and 5. The Enjoyment subscale consisted of factors 6,7 (reversed-scored) and 8.

Intrinsic Motivation

Intrinsic motivation was measured using the IMI on a per session basis. I created two measures from this questionnaire: perceived autonomy and perceived competence. Each measure was created for each session by calculating the mean score of its component items. The perceived autonomy subscale consisted of items 1, 2 and 3 (reversed scored), and the perceived competency subscale consisted of items 4,5 and 6 (reversed scored).

Coefficients of Reaction Time Variation

Coefficients of RT variation quantify RT intra-individual variability with respect to mean RT. Evidence has suggested that changes in motivation can be reflected in RT variation [222,223], and this was weakly supported by the findings of Experiment 2. Coefficients of RT variation were calculated on a per participant, per session basis by dividing the participant's standard deviation (SD) of non-stop trial RTs by their mean non-stop trial RT.

Stop-Signal Reaction Times

I calculated SSRTs for each session separately, excluding sessions where the assumptions of the race model did not hold (Section 5.3.4). For sessions that did meet the assumptions of the race model, I calculated $SSRT_{med}$ as described in [220]. The formula for calculating $SSRT_{med}$ is as follows:

$$\mathrm{SSRT_{med}} = Go\ Reaction\ Time_{median} - Stop\ Signal\ Delay_{med}$$

Stop-Signal $Delay_{med}$ (SSD) was calculated for each session using weighted least squares linear regression to predict SSD based on the probability of responding given a stop-signal. This was then used to estimate the SSD where the probability of the participant failing to inhibit was 50%.

6.3.5 Statistical Analysis

I did not assess the evidence for an effect of time in the analyses below because the order in which the task variants were presented to participants was counterbalanced within cohorts.

In all analyses, where appropriate, differences between groups were assessed with *post-hoc t*-tests. Where there was no evidence of a difference between group-means, I used Bayesian *t*-tests to assess the evidence for equality (3.3.5). Where Bayesian *t*-tests were between dependant groups, I used **paired** Bayesian *t*-tests.

Amount of Engagement

Differences in the amount of engagement between the task variants and the reimbursement schemes were assessed using a 2×3 mixed ANOVA of amount of engagement with reimbursement-scheme (flat-rate, pay-per-block) as a between-subjects factor and task variant (non-game, points, theme) as a within-subjects factor. I also plotted the amount of engagement across variants and reimbursement schemes to visually inspect differences.

Quality of Engagement

To assess the effect of the different task variants and reimbursement schemes on quality of engagement, I used a 2×3 mixed MANOVA of DBCI Engagement Scale subscale score (Interest, Attention and Enjoyment) with reimbursement-scheme (flat-rate, pay-per-block) as a between subject's factor and task variant (non-game, points, theme) as a within-subjects factor. I plotted bar charts of subscale scores separately by task variant to visually inspect differences.

Furthermore, to assess the strength of association between the DBCI subscales and amount of engagement I used a Pearson product-moment correlation matrix, combining across task variants and reimbursement scheme, with evidence quantified using BFs. Given my prior

hypothesis of a positive correlation, this was used as the alternative hypothesis in the Bayes Factor calculation.

Intrinsic Motivation

I assessed whether there was any evidence that the SDT concepts of autonomy or competency were being promoted by the theme or points variants respectively, and whether this differed across reimbursement schemes. I used a 2 × 3 mixed MANOVA of the perceived competency and perceived autonomy subscales of the IMI with reimbursement-scheme (flat-rate, pay-per-block) as a between subject's factor and task variant (non-game, points, theme) as a within-subjects factor. I plotted bar charts of subscale scores separately by task variant to visually inspect differences.

Coefficients of Reaction Time Variation

I assessed differences in coefficient of variation using a mixed ANOVA with reimbursement-scheme (flat-rate, pay-per-block) as a between subject's factor and task variant (non-game, points, theme) as a within-subjects factor.

Stop-Signal Reaction Times

To assess the effect of gamification on cognitive data I used a mixed ANOVA of SSRT with reimbursement-scheme (flat-rate, pay-per-block) as a between subject's factor and task variant (non-game, points, theme) as a within-subjects factor. I used box and whisker plots to compare SSRTs across the task variants.

6.3.6 Sample Size Determination

I based my sample size calculation on the findings of Prins and colleagues [50] where it was found that participants using a gamified cognitive task completed 48% more trials than those participants in the non-gamified control (*partial* η^2=.38, or d = 1.5). To detect a more conservative difference of d = 1.1, with a 5% alpha and 95% power, a sample size of 46 per cohort was required. I aimed to recruit 48 participants per cohort to allow for equal group sizes, giving a total sample size of 96.

6.4 Results

The data that form the basis of these results are available from the University of Bristol Research Data Repository (doi: 10.5523/bris.2l8sjofwxs7ha28v9pl27p6zk4).

6.4.1 Characteristics of Participants

Participants were recruited in January 2018, in two waves one week apart. Participants in the first wave formed the pay-per-block cohort, and participants in the second wave formed the

flat-rate cohort. In both waves, the intended sample size was met within hours of the study being posted on Prolific. A total of 118 participants signed up to take part in the study, and 106 (89.8%) completed at least one session. Where participants dropped out due to technical difficulties or other reasons, I continued to recruit in order to offset attrition. In the end, 90 participants (76.3%) completed the three required sessions within a 5-day period.

Of these 90 participants, 43 were in the pay-per-block cohort, and 47 were in the flat-rate cohort. The cohorts were not randomly determined, so I tested them for characteristic differences [236] but found no clear evidence for any differences (Table 6.2). The most common browser used to complete the experiment was Google Chrome (n=69, 70%), with others including Firefox (n=13, 13%), Internet Explorer (n=6, 6%) and Safari (n=10, 10%).

Table 6.2 Participants' demographic information, shown separately by task variant.

Demographic	**Pay-per-block cohort**	**Flat-Rate cohort**	**Test for difference**
Mean Age (SD)	32 (9.0)	35 (11.6)	t_{71}=1.197; p=.24
Number Male	15 (35%)	15 (32%)	$\chi^2_{1,73}$=.173; p=.68
Mean video game hours per week (SD)	8.6 (12.9)	4.9 (5.9)	t_{71}=1.928; p=.058
Median level of education	A-levels/Higher Education	Bachelor's Degree/University	$\chi^2_{4,73}$=9.345; p=.053
Mode ethnicity	41 (95%) White	43 (91%) White	$\chi^2_{3,73}$=1.045; p=.79

6.4.2 Amount of Engagement

Figure 6.4 shows the amount of participant engagement with the task variants across the two reimbursement schemes. I report Greenhouse-Geisser corrected values where Mauchly's test indicated sphericity to be violated. A mixed ANOVA showed no evidence for an effect of task variant on amount of engagement ($F_{1.9,163.8}$=.178, p=.82, *partial* η^2=.002) but there was strong evidence for a large effect of reimbursement scheme ($F_{1,88}$=185.8, p<.001, *partial* η^2=.679). The distribution of my amount of engagement variable was bimodal: skewed low in the flat-rate cohort and skewed high in the pay-per-block cohort (Figure 6.5). Given the similarity in amount of engagement between the task variants, I used Bayesian *t*-tests to assess equality. Combining across reimbursement scheme, I found substantial evidence that the amount of engagement was equal between the non-game and points variant (Bayes Factor (BF)=.13), the non-game and theme variant (BF=.12) and the points and theme variant (BF=.13).

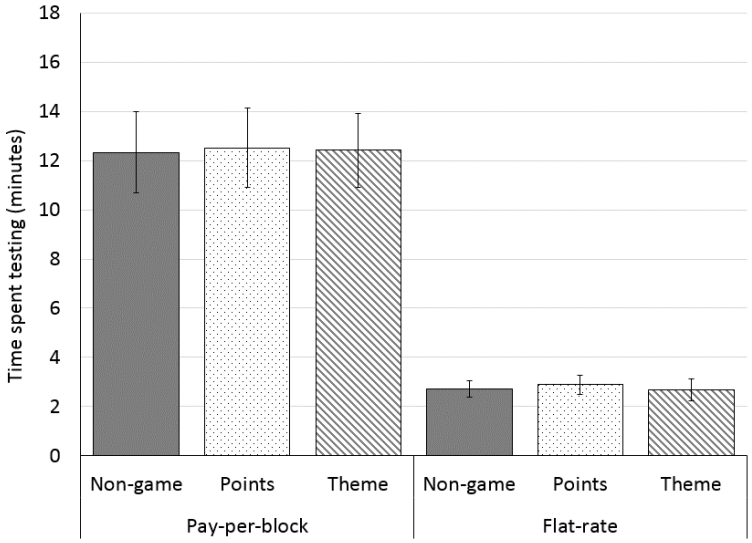


Figure 6.4 Mean number of minutes spent testing in each variant, shown separately by reimbursement scheme. Error bars represent 95% CIs.

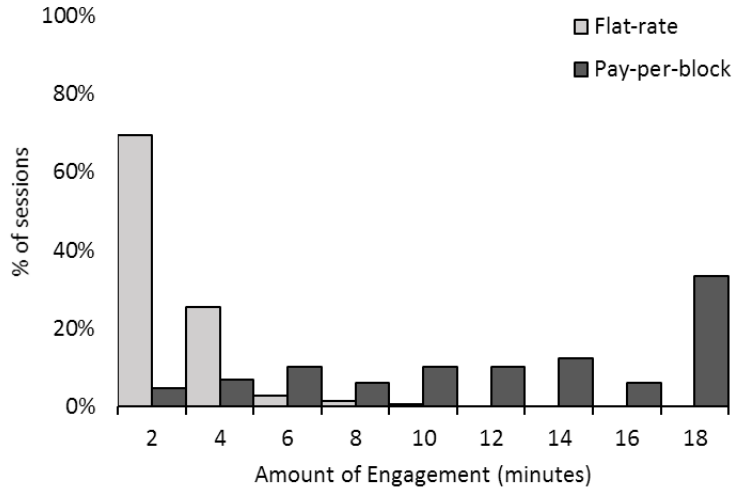


Figure 6.5 Distributions of amount of engagement in each reimbursement scheme. Combined across task variants. The y-axis represents the percentage of all the sessions in that cohort. I.e. the amount of engagement was only 2 minutes in nearly 70% of sessions completed by the flat-rate cohort.

6.4.3 Quality of Engagement

A mixed MANOVA was used to assess the effect of task variant and reimbursement scheme on DBCI Engagement Scale subscale score. There was strong evidence for large effects of task variant ($F_{6,83}$=3.913, p=.002, *partial* η^2=.115) and reimbursement scheme ($F_{3,86}$=3.730, p=.014, *partial* η^2=.115), but no evidence for an interaction between the two ($F_{6,83}$=1.326, p=.26, *partial* η^2=.087). Quality of Engagement scores were higher in the flat-rate cohort than the pay-per-block cohort.

Univariate ANOVA's showed strong evidence of medium effects of task variant on the Interest ($F_{2,176}$=10.831, p<.001, *partial* η^2=.110) and Enjoyment ($F_{2,176}$=6.316, p=.002, *partial* η^2=.067) subscales, but not the Attention subscale ($F_{2,176}$=.657, p=.52, *partial* η^2=.007). Figure 6.6 shows

the subscale scores separately across task variant and combined across reimbursement scheme.

Bayesian *t*-tests showed substantial evidence that the points and theme variants had equal Interest (BF=.12) and Enjoyment (BF=.12) scores. Attention subscale scores were equal in all three task variants: non-game and points (BF=.22), non-game and theme (BF=.13), and points and theme (BF=.16).

To investigate the associations between the three DBCI Engagement subscales and amount of engagement, I calculated correlation pairs between each subscale score and amount of engagement. I combined data over the three task variants and over reimbursement schemes, and assessed the evidence using BFs. BFs were calculated using 'positive correlation' as the alterative hypothesis, thus slightly weighting the evidence in favour of a positive correlation.

Overall, I saw substantial evidence that there was no association between amount of engagement and the Attention (*r*=-0.04, BF=.09) or Enjoyment subscales (*r*=-0.05, BF=.11). But there was substantial evidence of a small association between amount of engagement and the Interest subscale (*r*=0.16, BF=3.91).

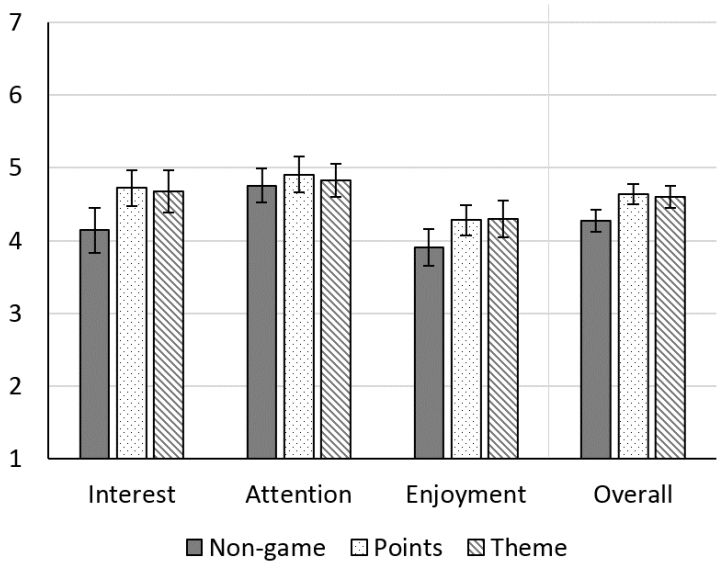


Figure 6.6 Mean subscale scores (and overall score) from the DBCI Engagement Scale, shown separately by task variant but combined across reimbursement schemes. Error bars represent 95% CIs.

6.4.4 Intrinsic Motivation

I used a mixed MANOVA to assess the effect of task variant and reimbursement scheme on perceived competency and perceived autonomy. I report Greenhouse-Geisser corrected values where Mauchly's test indicated sphericity to be violated. I saw strong evidence for a large multivariate effect of task variant ($F_{4,85}$=8.154, p<.001, *partial* η^2=.277) and a medium effect of

reimbursement scheme ($F_{2,87}$=4.625, p=.012, *partial* η^2=.096), but no evidence of an interaction between the two ($F_{4,85}$=1.003, p=.41, *partial* η^2=.045). Perceived autonomy and competency were higher in the pay-per-block cohort than the flat-rate cohort.

Univariate ANOVA's showed strong evidence of a large effect of task variant on perceived competency ($F_{1.68,148.11}$=23.282, p<.001, *partial* η^2=.209) but no evidence for an effect on perceived autonomy ($F_{2,176}$=.390, p=.68, *partial* η^2=.004). Figure 6.7 shows the scores for perceived autonomy and competency separately across task variant and combined across reimbursement schemes.

Paired post-hoc *t*-tests showed strong evidence that the points variant scored higher on perceived competency than either non-game (mean difference=.64, 95% CI .37-.91; t_{72}=4.678; p<.001; d=.55) or Theme (mean difference=.67, 95% CI .38-.96; t_{72}=4.542; p<.001; d=.53). Bayesian *t*-tests showed substantial evidence that the non-game and theme variants had equal perceived competency scores (BF=.14). Perceived autonomy scores were equal in all three task variants: non-game and points (BF=.14), non-game and theme (BF=.21), and points and theme (BF=.14).

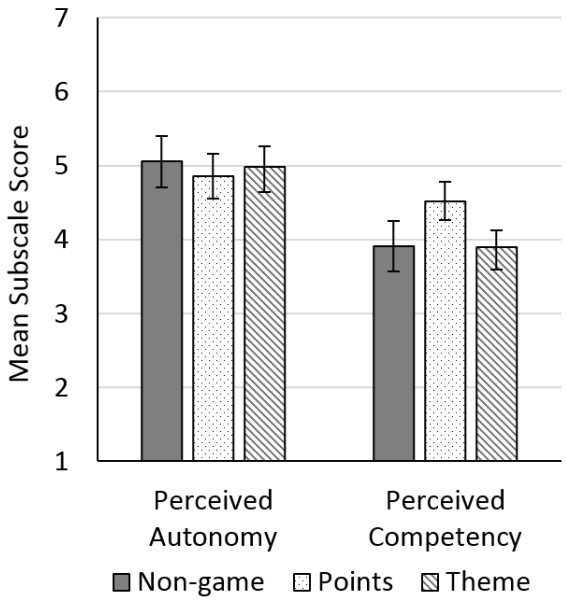


Figure 6.7 Mean subscale scores from the IMI, shown separately by task variant and combined across reimbursement schemes. Error bars represent 95% CIs.

6.4.5 Coefficients of Reaction Time Variation

A total of 17 participants were excluded from the cognitive data analysis for reasons prespecified in the protocol: 12 were excluded because their SST data failed to meet the assumptions of the race model in at least one task variant. 5 participants were excluded because their sorting accuracy was more than 4x the interquartile range away from the group

mean. As a result, the following analyses were performed on data from 73 participants, 37 from the pay-per-block cohort and 36 from the flat-rate cohort.

I analysed coefficients of RT variation across the task variants and reimbursement schemes. A mixed ANOVA of coefficients of variation showed weak evidence for a medium effect of task variant ($F_{2,140}$=4.504, p=.013, *partial* η^2=.060) on participants' RT variability, with lower coefficients indicating that there was less variability (Table 6.3). I saw no evidence for an effect of reimbursement scheme ($F_{1,70}$=.497, *p=.50*, *partial* η^2=.007).

Table 6.3 Mean coefficients of RT variation combined over reimbursement scheme, shown separately by task variant.

Variant	**Mean coefficients of variation (95% CI)**	**Mean intra-individual SD (95% CI)**	**Mean non-stop RT (95% CI)**
Nongame	18.9% (18.0-19.9)	110ms (105-115)	587ms (569-605)
Points	19.0% (18.1-19.9)	109ms (104-113)	580ms (562-599)
Theme	17.7% (16.9-18.4)	108ms (103-113)	612ms (594-630)

6.4.6 Stop-Signal Reaction Times

I analysed SSRTs across task variants and reimbursement schemes. A mixed ANOVA of SSRT showed strong evidence for a large effect of task variant ($F_{2,142}$=9.615, *p<.001*, *partial* η^2=.113) on participants' SSRTs. I also saw evidence for a large effect of reimbursement scheme ($F_{1,71}$=8.810, *p=.004*, *partial* η^2=.110), and weak evidence for a medium interaction effect ($F_{2,142}$=4.243, *p=.016*, *partial* η^2=.050). Mean SSRTs were higher in the flat-rate cohort. Mean SSRTs are shown in Table 6.4, and SSRTs across task variants are shown in Figure 6.8.

Table 6.4 Mean SSRTs, shown separately by task variant and reimbursement scheme

	Mean SSRT (95% CI)	
Task Variant	*Pay-per-block*	*Flat-rate*
Non-Game	301ms (287 - 314)	351ms (331 - 370)
Points	305ms (286 - 324)	341ms (322 - 360)
Theme	343ms (325 - 360)	355ms (334 - 376)

Combining across reimbursement scheme, I found substantial evidence that SSRTs were equal between the non-game and points variants (BF=.139), and very strong evidence for a difference between the non-game and theme variants (BF=41.36) and the points and theme variants (BF=38.16).

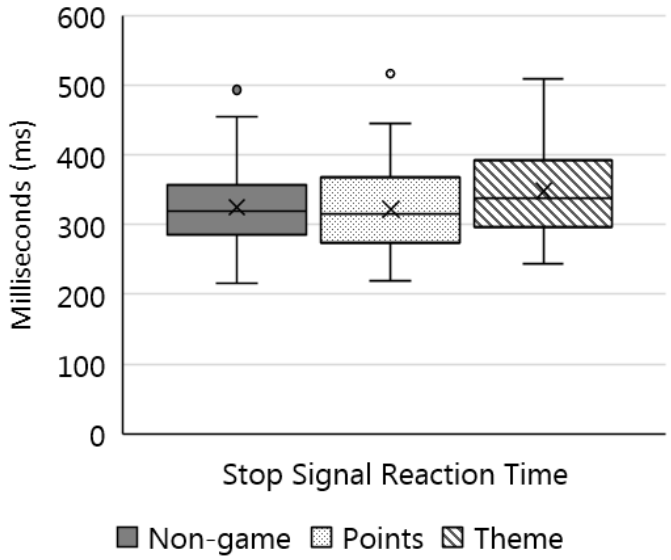


Figure 6.8 Boxplots of SSRTs, shown separately by task variant and combined over reimbursement scheme.

Exploratory analyses of secondary cognitive measures (non-stop RT, failed-stop RT, non-stop accuracy and stop accuracy) are presented in Appendix O (pg 162).

6.5 Discussion

6.5.1 Amount of engagement

Contrary to my hypothesis, I saw no effect of gamification on amount of participant engagement. Instead, I found substantial evidence that engagement with each task variant was equal, indicating that the gamification I employed was not capable of motivating participants to test for longer.

To the best of my knowledge, only three studies have directly assessed the effect of gamification on amount of engagement with a cognitive task. These studies compared non-gamified cognitive training tasks against gamified counterparts, and, contrary to my findings, found that participants in the gamified conditions completed more voluntary trials [50,52,232]. Potential explanations for this disparity in findings are: firstly, the studies by Prins [50], Dörrenbächer [52] and Boendermaker [232] were conducted on adolescents and children, who may be more attracted to video-game like tasks or more easily bored, thus enhancing the effect of gamification. Secondly, although participants in these studies *chose* to continue testing, choosing not to continue testing did not mean the participant could leave the session early, and this may have exaggerated the effect of gamification. Finally, these studies were all cognitive *training* studies, which made use of a game-shell to deliver their training (Step 4 gamification [198]), and this richer form of gamification may have been more motivating. Overall, it seems that further research is necessary to establish the contexts in which gamification can motivate additional engagement.

6.5.2 Financial Incentive

This experiment provided strong evidence for a large effect of reimbursement scheme on amount of engagement. This was in line with my hypothesis and was unsurprising given that financial incentive is known to be the most important factor in crowdsourced-worker's motivation [237]. Participants in the pay-per-block cohort tested for, on average, six times longer than those in the flat-rate cohort.

In both cohorts, the distribution of amount of engagement was skewed (Figure 6.5). I acknowledge that better piloting of participant rewards might have increased the sensitivity of my measure and revealed an effect of gamification, potentially by offering a lower payment per block. However, Prolific enforces a minimum payment of £5 per hour for participant rewards and this study tested that limit. It is worth considering the possibility that the motivational effect of financial incentive might be so large that it masks any possible effect of gamification. If so, any effects of gamification on engagement would only be apparent when participant reward was very low.

One quarter of the pay-per-block cohort tested for the maximum amount of time in every task variant, despite being paid only 3p for the final block of the task. This is a strikingly low rate of pay. Possible explanations for this behaviour include: the gradual reduction in incentive may have made it difficult for participants to decide when to drop out. The heightened focus on extrinsic reward in crowdsourced-workers may have made it difficult for them to resist the offer of additional reward, regardless of how low it was. Finally, competition over study places on Prolific is quite fierce: perhaps participants saw risk in quitting before all possible payment was acquired because there was no guarantee of other available work. Regardless, this finding suggests that merely offering an incentive, regardless of how small, can be enough to increase the amount of engagement. However, there is growing evidence that good participant performance necessitates fair pay [238,239], and the ethics of paying participants less than the minimum wage must also be strongly considered [240,241].

6.5.3 Quality of engagement

Despite no evidence that gamification could increase amount of engagement, my data indicate a robust, yet small, positive effect of gamification on quality of engagement. Both the points and theme variants were rated as equally interesting, attention-grabbing and enjoyable, and both were rated as being more enjoyable and interesting than the non-game variant. The points variant was not rated more highly than the theme variant, as was found in Experiments 1 and 2. This may be because the within-subjects design allowed for comparison between the task variants. Previously I have speculated that the theme variant is disappointing when

considered alone, since “it looks like a game but doesn’t play like one”. When compared against the non-game variant however, it is not surprising to see it rated as more enjoyable and more interesting. Finding both types of gamification to improve quality of engagement is in line with the rest of the literature (Section 2.4.6).

I hoped the IMI would explain the 'motivational mechanisms influencing participants’ quality of engagement, however it provides only a partial picture. In line with my hypothesis, the points variant provided a greater sense of competency than the other two variants. The scoring system inherent to the points variant provides constant feedback to the participant, and SDT suggests that feedback on performance helps to meet the competency need [134,151]. This explains the points variant’s quality of engagement score.

The theme variant provided the same quality of engagement as the points variant, but contrary to my hypothesis, this was not explained through participants’ greater sense of autonomy. It has been argued that graphical customisation, such as player avatars, *are* sufficient to promote autonomy [242,243], and giving the participant the ability to choose the graphical theme of the next level was intended to do just that. Nevertheless, despite best intentions, changing the task’s background-graphics did not change anything fundamental about the task itself and as such may have failed to provide meaningful autonomy. Alternatively, the presence of extrinsic motivation (in the form of financial incentive) may have blunted any differentiation of autonomy between the task variants [226] (i.e., participant’s autonomy was being driven by their choice to work for money, and not by their subjective experience of the task).

I hypothesised a positive relationship between amount and quality of engagement, but I found substantial evidence that there was no association between the two: at least on the Attention and Enjoyment subscales of the DBCI Engagement Scale. I did find evidence of a weak association between the Interest subscale and the amount of engagement, comparable in size to the association found in Experiment 2. I acknowledge that this is weak evidence, but the small correlation may be the result of the (unideal) bimodal distribution of amount of engagement. This association provides a glimmer of evidence that amount and quality of engagement are related, as theorised by my conceptualisation of engagement (2.5.1).

6.5.4 Cognitive Data

I found clear evidence that that SSRTs in the theme variant were ~40ms longer than in the non-game and points variants. Given the within-subjects design of this study, this indicates that the theme variant does cause response slowing. Given the increased visual complexity of

the theme variant compared to the other variants, one possibility is that this slowing happens on a perceptual level, rather than on a cognitive level. In contrast, I saw substantial evidence that SSRTs from the non-game and points variants were equal, corroborating the findings of Experiments 1 and 2: that points do not impact the cognitive data collected by a task.

SSRT estimates become more accurate as more blocks are completed, as it takes the staircases several blocks to converge to a 50% failed-inhibition rate. Given the low amount of engagement in the flat-rate cohort, the large effect of reimbursement scheme on SSRTs is almost certainly the result of overestimation during the SSRT calculation process.

I investigated the potential of coefficients of RT variation as a behavioural measure of engagement. My hypothesis was that more engaging task variants would result in lower RT variation. Just as in Experiment 2, my data show the theme variant to have the lowest coefficient of variation, and therefore the lowest amount of intra-individual variability: potentially indicating heightened participant engagement. However, this study's within-subjects design reveals a flaw in the use of coefficients of RT variation for quantifying engagement. Table 6.3 shows that intra-individual standard deviations were very comparable between the task variants (~110ms), but that RTs in the theme variant were ~30ms longer. I posit the lowered coefficient of variation in the theme variant is an artefact of these slowed response times, rather than a result of increased focus. Other researchers have sought behavioural correlates of heightened engagement [49], but it appears that coefficients of RT variation are unsuitable.

6.5.5 Limitations

I acknowledge several limitations which affect the generalisability of these findings. Firstly, the bimodal distribution of amount of engagement likely reduced the sensitivity of my measure and weakened the power of my statistical tests. Secondly, the effect of financial incentive may be exaggerated in this online crowdsourcing population compared to the general population, as financial payoff is known to be the most important factor in why a user does crowdsourcing work [237]. Thirdly, the gamification I used was superficial. Both gamified tasks were presented as a game, but the fundamental nature of the task were no different to the non-game control. This likely limited any potential effect of gamification.

6.6 Chapter Summary

This chapter described Experiment 3, which investigated the effects of gamification and financial incentive on amount of engagement. I used a mixed design, with two cohorts of participants completing all three task variants (non-game, points, theme) over a five-day

period. I measured amount of engagement using *ad-libitum* test time. Each test session was of variable duration, with the SST being paused every two minutes to ask participants if they wished to continue testing. To investigate the effect of financial incentive, the two cohorts used different reimbursement schemes: one cohort received payment of £1.25 for completing each session regardless of time spent, while the other was incentivised to complete each block in exchange for a linearly decreasing financial reward. I also measured quality of engagement using the DBCI Engagement scale, and intrinsic motivation using the IMI.

Contrary to my hypothesis, I saw no evidence that gamification could increase amount of engagement. Instead, I found strong evidence that amount of engagement was equal between the three task variants, regardless of reimbursement scheme. There was a large effect of financial incentive, with participants in the flat rate cohort typically completing the minimum amount of testing possible. The results of this study indicate that financial incentive has a much larger effect than gamification on amount of engagement. The DBCI Engagement Scale indicated medium effects of task variant on participant quality of engagement, with the points and theme variants being rated as more interesting and more enjoyable than the non-game variant.

Overall, the findings of this study indicate that introducing points to a SST does not impact the cognitive data collected, and improves quality of engagement. Introducing graphical theme is detrimental to cognitive measures, but similarly improves participant experience. Neither gamified task variant had any impact on the amount of time which participants chose to test for, but incremental financial incentive appears to be an effective way of maintaining engagement. In the next and final chapter, I synthesise my findings from Experiments 1,2 and 3, assessing my results in the context of the wider literature.

Chapter 7: General Discussion

7.1 Chapter Aims

This thesis describes three years of research into the suitability of gamification as a tool for increasing engagement with gamified cognitive tasks. I have conducted a systematic review of the field, developed my own online platform for deploying gamified cognitive tests, and conducted three experiments into the effects of gamification on participant engagement and cognitive measures. This chapter synthesises the evidence I have collected and reflects on my findings in the context of the wider literature. It has three aims, namely to:

1. Consolidate my findings to answer and contextualise the central question of my thesis: *Is gamification a suitable tool for increasing participant engagement with cognitive tests?*
2. Address the limitations of my research
3. Discuss the challenges remaining in the field and suggest directions for future research

7.1.1 Is gamification a suitable tool for increasing participant engagement with cognitive tests?

7.1.2 Does the gamification of a cognitive test affect the data collected?

Evidence from my systematic review (Section 2.4.6) indicates it *is* possible to design gamified tests that collect valid cognitive measures [45,47,97,100,103,139]. Similarly, Bayesian *t*-tests from Experiments 1,2 and 3 provide substantial evidence that introducing points to a test does not affect RT, accuracy, or even compound measures such as stop signal reaction time (SSRT). The within-subjects design of Experiment 3 provides particularly strong evidence for this, with equality of SSRTs between the non-game and points variants.

That said, a handful of studies in my review suggested a negative impact of gamification on cognitive data. Some found that their gamified output measures correlated with more cognitive measures than intended, indicating they weren't indexing a single cognitive function [51,89]. Others saw lengthened RTs [49] or reduced training effect [101]. My studies consistently showed deleterious effects of themed graphics on cognitive measures. In Experiment 1 the theme variant caused large increases in RT and reductions in no-go accuracy. In retrospect this isn't surprising given the complexity of the stimuli, but despite simplifying and colour matching stimuli between the non-game and theme variants in Experiments 2 and 3, the theme variant still caused longer SSRTs. Overall, my data suggest that themed graphics cause response slowing and lower response accuracy, and I would caution against the use of graphics in cognitive tests sensitive to RT or where response accuracy is of importance.

There is evidence that poor participant motivation can negatively affect data quality [21–23]. It has been suggested that the motivational effect of gamification might reverse this [47]: improving participant performance by reducing within-subject variability, increasing accuracy and speeding response times. Neither my systematic review nor my empirical research found evidence to support this idea, although some other studies have indicated that points can increase participant performance on tasks such as image tagging [154], working memory training [32] and maths tests [230]. It is debatable whether psychological researchers should even want gamification to improve performance. On one hand, increasing motivation and improving performance might allow a truer measure of an individual's maximum cognitive capability. On the other hand, improved performance might result in a measure that does not reflect the daily cognitive functioning of that individual, threatening test validity and undermining diagnostic potential [93].

In summary, my findings indicate it is possible to carefully introduce game elements (i.e., points) to a cognitive test without having a measurable impact on the data collected. However, certain game design elements (i.e., graphics) may introduce additional cognitive load and negatively impact the test's functioning.

7.1.3 Does the gamification of a cognitive test affect participants' quality of engagement?

Studies in my review unanimously asserted that gamification was effective for increasing participants' quality of engagement (Section 2.4.6). My findings were not so clear cut, with points and theme having different effects on quality of engagement, and inconsistent participant ratings of theme.

In Experiments 1 and 2, the points variant provided the highest quality of engagement of the three task variants. In Experiment 3, the points variant was rated as being more enjoyable and interesting than the non-game variant, but equal to the theme variant. In all studies, while there was consistent evidence for a difference in participant ratings between the non-game and the points variants, this difference was not very large (typically ~5%). The small size of this effect was reinforced by Experiment 3 where participants could compare their experiences across the three task variants, and yet the difference in scores remained modest. Overall, my evidence suggests a robust, yet small, positive effect of points on quality of engagement.

In Experiment 3 I used the Intrinsic Motivation Inventory to explore the motivational mechanisms behind participants' quality of engagement. I found strong evidence that the competency need was being met by the points variant and consider it likely that the

satisfaction of this psychological need is causally connected to the heightened quality of engagement it provided. While intuitive, it is important to note that points do not *always* induce intrinsic motivation, for example, Mekler and colleagues have conducted studies where points neither harmed nor fostered intrinsic motivation [154,212]. It has been suggested that the same game element might induce intrinsic or extrinsic motivation (or no motivation) depending on the context and the individual [38,226,227].

The effect of theme on quality of engagement was variable. Experiment 1 saw the theme variant rated second to points, with no evidence it was any more engaging than the non-game variant. In Experiment 2, a similar theme was rated as less engaging than the non-game variant. In Experiment 3, the points and theme variants were rated as being equally interesting, attention grabbing and enjoyable; and both more interesting and enjoyable than the non-game variant.

The attraction of narrative framing and graphics is more subjective than points [244], and it's possible the themes used were unappealing to some participants. Adverse effects of gamification on quality of engagement have been found before, potentially as a result of expectation violation i.e., disappointment that the task isn't as fun as it looks [225] (Section 5.5.1). It is also possible that participants found the theme variants more difficult: they took longer to respond to stimuli but the response windows were no longer than in the non-game or points variants. A perceived increase in difficulty may have lowered participants' quality of engagement.

In Experiment 3, the IMI did not explain the motivational effects of the theme variant. My hypothesis was that allowing participants a choice of 'destination' would create a sense of autonomy [242]. It has also been argued that stories are important for task meaningfulness [37,136]. They allow participants to see their own actions within a context, thus giving them an illusion of meaningfulness and autonomy. However, the IMI indicated that the theme variant did not provide any more autonomy (or competency) than the points or non-game variants. Experiment 3 showed the theme variant to provide quality of engagement equal to the points variant, despite not satisfying autonomy or competency. This suggests that when the theme variant improves quality of engagement, it does so via a mechanism not captured by the IMI.

In summary, my data suggest that adding points to a cognitive test helps to satisfy participants' competency needs and raises their quality of engagement, though the effect is limited. Themed graphics and narrative may also improve quality of engagement, though further research is needed to establish both this and the motivational mechanisms of action. In

general, it seems that gamification has the potential to improve quality of engagement, but more research is needed to understand how we might increase the size of this effect.

7.1.4 Does the gamification of a cognitive test affect participants' amount of engagement?

The effects of gamification on amount of engagement with cognitive tests are decidedly under-researched. To the best of my knowledge, three studies have directly looked at this topic (technically these were all cognitive training tasks rather than tests) [50,52,232]. All found evidence in gamification's favour, but they targeted children and used different measures of engagement to my own (see Section 6.5.1 for more details). This raises questions about the generalisability of their findings to a less specific population.

On the basis of my empirical results, the answer to the above question is no. Experiments 2 and 3 measured amount of engagement in two different ways (over days and over minutes), and my data clearly indicate that neither points nor theme have any effect on participants' amount of engagement with a cognitive test. Despite some evidence that quality and amount of engagement are weakly associated, the small effects of gamification on quality of engagement did not translate into increased amount of engagement.

7.1.5 Synthesising the findings

Contrary to my expectations, the combined evidence leads me to conclude that gamification is not a suitable tool for increasing engagement with cognitive tests. It is possible to gamify a task using points while not invalidating cognitive measures, but the effect on quality of engagement is small, and there is no impact on amount of engagement. Themed graphics have a less reliable effect on quality of engagement and there is strong evidence they cause response slowing. Again, there is no effect on amount of engagement.

When this project began, gamification was heralded as a solution to all engagement problems [87]. The idea that including game-like elements in a boring task might make it more enjoyable was intuitive, and the application of gamification *should* have been easy. But gamification research has matured over the last few years [245], and reviews from the wider health sciences literature have struggled to show clear evidence of effectiveness.

Brown and colleagues reviewed the impact of gamification on adherence to online mental health interventions [42]. They found that while gamification had been widely applied in this context, there were no studies explicitly assessing the effect of gamification on adherence. Comparing between studies, they saw no evidence that adherence was higher in studies using gamification compared to those that did not, nor was there any relationship between the

number of game design elements included and adherence. Johnson and colleagues' [246] review of gamification for health and wellbeing interventions concluded that gamification could have a positive impact, but they found the evidence to be of moderate to low quality. Only 59% of studies in their review reported positive findings [247], with the rest reporting mixed or null effects. Ambiguous effects were particularly common in interventions designed to address cognition. They concluded that better study designs were needed isolate the impact of gamification on interventions, and that the state of evidence was such that "little can be said conclusively". Sardi and colleagues' [71] similar review of gamification in e-health found a shortage of empirical evidence: with only half the papers that met their inclusion criteria providing any evidence whatsoever. Again, they found positive reports of gamification's effectiveness, but only with respect to short-term engagement driven by real-world extrinsic rewards. Taken together, these reviews suggest that evidence is rare in favour of gamification's effectiveness in the health sciences.

It has gradually become apparent that successful gamification is more difficult to deliver than was assumed several years ago, and I consider my negative conclusions to fit with the current state of the field. That said, I do not expect that gamification will *never* be a suitable tool for increasing engagement with cognitive tasks, rather that more research is needed.

7.2 Limitations

7.2.1 Paid crowdsourced samples for engagement research

Most participants I studied were drawn from two online pools of crowdsourced workers: MTurk and Prolific. The results of Experiment 1, where I compared a laboratory group against an MTurk group (Section 3.5.1), indicated that online samples could provide comparable results to laboratory samples. This is supported by the literature, which also highlights how the demographics of crowdsourced samples better reflect the general population than typical undergraduate lab samples [206,248,249].

However, members of these online samples are unusual. They have chosen to sign up to a crowdsourcing website to complete surveys, microtasks and research studies in exchange for money. On Prolific, participants can select which studies they wish to take part in on the basis of reward, task length and task content. Participants in these online samples enter the study with their own goals in mind, whether they be financial or prosocial [237,250]. These factors serve to set crowdsourced online samples apart from the general population and potentially limit the generalisability of my findings. But with respect to my research, crowdsourced workers *are* a valid population to study. The popularity of online health research continues to

grow, meaning many research findings are based on these unusual populations. Accordingly, we need to study both their representativeness and methods of engaging them.

I was surprised to find little effect of gamification on amount of engagement in my studies. However, it is possible that crowdsourced workers are a population unsuited to engagement research, for two reasons. Firstly, previous research has shown that financial recompense is the primary motivator for crowdsourced workers [237,251], and job listings on both MTurk and Prolific are heavily focussed on hourly rate of pay. A typical worker comes to these sites looking to make money; and the purpose or subjective experience of the task is of secondary importance. This would explain gamification's small effect on engagement: participants' motivation was already saturated by financial reward alone. Secondly, as I've mentioned several times, there is substantial evidence that engagement-contingent extrinsic rewards (as were used in my studies) undermine intrinsic motivation [226], thereby hindering gamification's theorised mechanism of action. As such, I acknowledge that paid crowdsourcing work-marketplaces such as MTurk and Prolific may be unideal for studying the effects of gamification on engagement.

A workaround would be to use unpaid crowdsourcing (such as the Great Brain Experiment did: www.thegreatbrainexperiment.com) to study engagement, thus avoiding confounding due to financial incentives. A recent review has suggested that gamification is very effective in this environment: increasing quantity and quality of work [252]. The use of narrative to imbue tasks with meaning may also improve quality of work in this context [250,253]. Also, recent evidence indicates that intrinsic motivation can rebound once extrinsic rewards are removed [254,255]. This suggests that financial incentive could be used to acquire an initial sample, but rewards could be gradually phased out, leaving any underlying intrinsic motivation remaining for study. However, if a researcher hopes to maintain a meaningful sample size for a long period of time without constant financial reward, they're under more pressure to develop high quality and intrinsically motivating gamification.

7.2.2 Superficial gamification

The gamification I employed in my experiments was relatively superficial (Appendix J - pg146). Though the gamified task variants appeared gamelike, they did not change the fundamental nature of the task. I decided to use this superficial level of gamification for three reasons. Firstly, this level of gamification is common in the health sciences, with many gamified interventions or tasks using only one game element [42], and there is some evidence that superficial gamification can be effective [256–258]. Secondly, I was concerned about

invalidating cognitive measures. My systematic review provided some examples of misjudged gamification worsening participant performance [49,89,101], and so I kept my gamified task variants as close to their non-game counterparts as possible to preserve comparability of data. Thirdly, to understand the effects of game design elements on motivation and cognition, I needed to investigate them individually even though a single game design element cannot provide a rich experience. I acknowledge that the restricted level of gamification I employed may have limited any potential effects on engagement.

In the wider world, the most successful examples of gamification make use of rich narrative content, high fidelity graphics and a complementary blend of game design elements: e.g., Zombies Run (zombiesrungame.com), Pokemon Go (www.pokemongo.com), FoldIt (fold.it) etc. Furthermore, the 'golden rule of games design' is to personalise your gamification to your intended audience [221], and to work with them when designing your games [64,66,259]. My intended audience was crowdsourced workers, a very disparate group, and I did not conduct user centred design nor include them in a co-creation process. I acknowledge that this lack of user-led approach may have resulted in gamified task variants which did not appeal to my target audience, limiting any potential effects of gamification on engagement.

7.2.3 Questionnaire measures of engagement

I designed the questionnaire of quality of engagement used in Experiments 1 and 2 based on items used by Hawkins and colleagues and Miranda and Palmer [47,49]. Ad-hoc questionnaire creation is common in the field (Section 2.4.5). Nevertheless, I acknowledge that had I used a more established measure of quality of engagement, such as the Player Experience Needs Satisfaction scale (PENS) [260] (see 7.3.2), my measure might have been more sensitive to differences in subjective experience between the task variants.

In Experiment 3 I switched to the DBCI Engagement scale. Primarily because it was constructed with more rigour than my own measure, but also to contribute evidence towards its validation. However, I acknowledge that this limits the comparability of quality of engagement scores between my three studies.

7.3 Challenges and Future Work

7.3.1 Building a better foundation for gamification research

I suggest that the future of gamification research will involve two parallel streams of study: the identification of the salient features of game design elements and the study of their effects, and the gradual construction of a game design element taxonomy.

Game design elements are subjective and difficult to describe precisely [38]. Attempts have been made at categorising game design elements, design patterns, models of play, etc, into frameworks of gamification [62,261,262] (for a review see [263]). However, these frameworks have failed because the terms they use are often ambiguous and implementation details are vague [38,264], making them difficult to validate empirically. This arises because we do not have a good empirical or theoretical understanding of the salient features (or 'active ingredients') of game design elements, and accordingly we cannot precisely describe them.

My research approach was to study game design elements individually: investigating their effects on motivation and task performance: but it did not go deep enough. We need even lower level, granular research to unpick what aspects of a game design element influence its effects [227]. What is it about points (for example) that improves quality of engagement: is it the provision of constant feedback? The thrill of beating your high score? The sense of reward? Or simply the 'juicy' way in which they are presented [221]? I posit that abstract game design elements, such as points, are actually combinations of even lower level game design elements, together describing the precise implementation of the higher-level game element.

As these 'atomic' game design elements are specified, we should categorise them in a taxonomy. An evidence-based example of which comes from the field of behaviour change [265]. The Behaviour Change Technique Taxonomy was constructed using the Delphi method [266] in combination with empirical evidence, and represents an expert consensus opinion on the 93 techniques used in behaviour change interventions. Widespread usage has been facilitated through online training courses (www.bct-taxonomy.com), and the taxonomy can be updated to reflect changes in evidence and theory. The formation of such a taxonomy would provide a consistent terminology with which to describe game design elements, facilitating the comparison of gamified tasks, easing replications, allowing us to estimate the generalisability of our findings and highlighting gaps in the literature in need of further research.

7.3.2 Improving measures of engagement

In Section 2.5.1 I defined engagement as a multidimensional construct, of which two dimensions were quality and amount of engagement. This definition was based on a review by Perski and colleagues [150]. However, the definition and conceptualisation of engagement is still intensely debated [149], and struggles seem particularly focussed on quantifying participant's subjective experience [267]. A wide variety of questionnaires have been used to examine participants' quality of engagement with cognitive tasks (Section 2.4.5), and several

studies (including my own) used ad-hoc questionnaires. There is also a proliferation of 'developed' questionnaires designed to measure similar concepts: the Intrinsic Motivation Inventory [235], the Gaming Motivation Scale [268], the Technology Acceptance Model [269], the User Engagement Scale [270], the Dispositional Flow Scale [129], the Gaming Engagement Questionnaire [267], the Immersive Experience Questionnaire [271], the Player Experience Needs Satisfaction Questionnaire (PENS) [260], and so on. It may be that one of these questionnaires is the *best* tool for capturing post-hoc subjective experience of an activity, but their sheer number means any evidence of validity is spread thin and limits the comparability of studies [272]. Furthermore, the use of proprietary questionnaires (such as the PENS and Dispositional Flow Scale) prevents sharing and openness of measures, thus limiting the evidence that can be accumulated for (or against) their validity. Future researchers should seek to establish a questionnaire measure of subjective experience as a gold standard in the field, as this will enormously increase comparability of studies.

Experiments 2 and 3 provided only weak evidence of an association between quality and amount of engagement. The bimodal distribution of amount of engagement recorded in Experiment 3 may have weakened the association between quality and amount of engagement, but Experiment 2 had no distributional problems and still showed only weak evidence for an association. However, it is possible that both of my measures of amount of engagement were too crude and thus insensitive to the effects of gamification. On large commercial websites (such as Facebook, or Amazon), vast amounts of user interaction data are aggregated and analysed to understand how and why users engage with the site. Tools such as Google Analytics and Amplitude allow for exploration of these rich datasets and presumably, given the success of Google and Facebook, they provide considerable insight into user engagement. To the best of my knowledge, 'analytics' has only made small inroads into academic engagement research [148,273], potentially because academics often do not have access to the large userbase required. Future researchers should consider embedding analytics tools into their studies of gamification to provide a more sensitive measures of, and insight into, user engagement.

Finally, though I've conceptualised engagement as quality and amount of engagement, my investigations into coefficients of RT variation and loss-of-focus events in Experiments 2 and 3 touched on a third dimension of engagement: immersion. This facet of engagement reflects a user becoming absorbed in a task and losing awareness of the outside world [271]. There are several promising paradigms for measuring immersion, such task switching [271], dual tasking

[274,275], and eye tracking [276]. Future researchers should not neglect the value of these objective measures for triangulating evidence on engagement.

7.3.3 Cognitive Taskification

In every gamified application there is a tension between gamification and functionality. These design limitations are particularly pertinent for gamified cognitive tests, where the smallest adjustment can cause substantial changes in outcome measures. In the theme variant of my stop-signal task I carefully implemented superficial gamification and yet saw detrimental effects on cognitive data which affected the validity of the test. Furthermore, it is unclear whether we will ever develop a strong enough understanding of gamification that we will be able to increase the amount of engagement with a cognitive test without invalidating the task.

Throughout my thesis, I have taken a bottom-up research approach: Adding game elements to existing cognitive tasks to make them more appealing. Future researchers might consider the reverse approach: taking an existing game and building cognitive measures into the backend, thus transforming it into a cognitive assessment. Hugely popular video games such as Overwatch, League of Legends, Minecraft and Hearthstone collect vast amounts of interaction data from their players, in a range of situations. League of Legends, for example, involves rapid decision making, fast responses to stimuli, acting on hidden information, remembering the 2D positions of hidden objects, and more. Many of these gameplay features have obvious cognitive-measure counterparts. Given the vast amount of game-data, machine learning might prove a useful tool: extracting counterparts to existing cognitive measures and providing insight into possible *new* measures. Using existing games for cognitive assessment in this way circumnavigates the need for scientists to develop their own games from scratch (which is both risky and expensive) and may allow the large-scale collection of population norms. But most importantly, it may provide psychologists with a genuinely engaging cognitive test.

7.4 Conclusion

This thesis describes three years of research into the suitability of gamification for increasing participant engagement with cognitive tests. Combining evidence across a systematic review and three experimental studies, I conclude that gamification, as it currently stands, is not the silver bullet for engagement that it was hoped to be.

Experiments 1,2 and 3 demonstrated that it was possible to use points to gamify a cognitive test without impacting the cognitive data collected. Points had a robust, yet modest, effect on participants' quality of engagement: facilitated by the satisfaction of their competency need. In contrast, the introduction of themed graphics and narrative to a cognitive test had

deleterious effects on cognitive data and mixed effects on participants' quality of engagement. Crucially, Experiments 2 and 3 suggest that gamification does not increase participants' amount of engagement.

Within the wider field of health sciences my findings are not out of place, with recent reviews finding unclear effects of gamification on engagement. At the beginning of this project I regarded such unclear findings with optimism. Now, having conducted several studies in the field, I can appreciate why researchers have struggled to utilise gamification in the health sciences. The limitations of my work are reflected in the field: rich gamification is difficult to implement without undermining the task's purpose, measures of engagement are not capturing the full picture, and accessing unconfounded research samples is an ever-present challenge.

Nevertheless, I believe that gamification will continue to improve in quality as both a tool and a field of research. Evidence from outside the health sciences, where gamification can be applied more liberally, is much more positive. Looyestyn and colleagues found clear evidence of positive effects on engagement with online education programmes [43], and Hamari and colleagues have found that gamification can increase motivation in a range of areas including work, innovation and education [39]. Much more research is required: we need a taxonomy of atomic game design elements to guide construction of an evidence base, and we have little understanding of *what* game design elements have *what* effects, in *which* contexts, with *whom*. But research methodologies are improving, gamification conferences are growing, and empirical evidence is beginning to accrue. We might yet see the day when gamification facilitates long term engagement with a cognitive monitoring program.

References

1. Crump MJC, McDonnell JV, Gureckis TM. Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research. PLOS ONE 2013 Mar 13;8(3):e57410. [doi: 10.1371/journal.pone.0057410]

2. Griffiths F, Lindenmeyer A, Powell J, Lowe P, Thorogood M. Why Are Health Care Interventions Delivered Over the Internet? A Systematic Review of the Published Literature. J Med Internet Res 2006 Jun 23;8(2). PMID:16867965

3. Peer E, Brandimarte L, Samat S, Acquisti A. Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. J Exp Soc Psychol 2017 May;70:153–163. [doi: 10.1016/j.jesp.2017.01.006]

4. Woods AT, Velasco C, Levitan CA, Wan X, Spence C. Conducting perception research over the internet: a tutorial review. PeerJ 2015;3:e1058. [doi: 10.7717/peerj.1058]

5. Bennett GG, Glasgow RE. The delivery of public health interventions via the Internet: actualizing their potential. Annu Rev Public Health 2009;30:273–292. PMID:19296777

6. Buhrmester M, Kwang T, Gosling SD. Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? Perspect Psychol Sci 2011 Jan 1;6(1):3–5. [doi: 10.1177/1745691610393980]

7. Shapiro DN, Chandler J, Mueller PA. Using Mechanical Turk to Study Clinical Populations. Clin Psychol Sci 2013 Apr 1;1(2):213–220. [doi: 10.1177/2167702612469015]

8. Birnbaum M. Human research and data collection via the Internet. Annu Rev Psychol 2004;55:803–832. [doi: 10.1146/annurev.psych.55.090902.141601]

9. Birnbaum M. Psychological experiments on the Internet. 1st ed. Elsevier; 2000. ISBN:978-0-12-099980-4

10. Etter J-F. Comparing the Efficacy of Two Internet-Based, Computer-Tailored Smoking Cessation Programs: A Randomized Trial. J Med Internet Res 2005;7(1):e2. [doi: 10.2196/jmir.7.1.e2]

11. Farvolden P, Denisoff E, Selby P, Bagby RM, Rudy L. Usage and Longitudinal Effectiveness of a Web-Based Self-Help Cognitive Behavioral Therapy Program for Panic Disorder. J Med Internet Res 2005;7(1):e7. [doi: 10.2196/jmir.7.1.e7]

12. Wangberg SC, Bergmo TS, Johnsen J-AK. Adherence in Internet-based interventions. Patient Prefer Adherence 2008 Feb 2;2:57–65. PMID:19920945

13. Kelders SM, Kok RN, Ossebaard HC, Van Gemert-Pijnen JE. Persuasive System Design Does Matter: A Systematic Review of Adherence to Web-Based Interventions. J Med Internet Res 2012 Nov 14;14(6). PMID:23151820

14. Dumville JC, Torgerson DJ, Hewitt CE. Reporting attrition in randomised controlled trials. BMJ 2006 Apr 22;332(7547):969–971. PMID:16627519

15. Zhou H, Fishbach A. The Pitfall of Experimenting on the Web: How Unattended Selective Attrition Leads to Surprising (Yet False) Research Conclusions. J Pers Soc Psychol 2016 Jun 13; [doi: 10.1037/pspa0000056]

16. Christensen H, Mackinnon A. The Law of Attrition Revisited. J Med Internet Res 2006;8(3):e20. [doi: 10.2196/jmir.8.3.e20]

17. Eysenbach G. The Law of Attrition. J Med Internet Res 2005 Mar 31;7(1):e11. [doi: 10.2196/jmir.7.1.e11]

18. Murray E, Khadjesari Z, White I, Kalaitzaki E, Godfrey C, McCambridge J, Thompson S, Wallace P. Methodological Challenges in Online Trials. J Med Internet Res 2009;11(2):e9. [doi: 10.2196/jmir.1052]

19. D'Angiulli A, LeBeau LS. On Boredom and Experimentation in Humans. Ethics Behav 2002 Apr 1;12(2):167–176. PMID:12956142

20. Corbett A, Owen A, Hampshire A, Grahn J, Stenton R, Dajani S, Burns A, Howard R, Williams N, Williams G, Ballard C. The Effect of an Online Cognitive Training Package in Healthy Older Adults: An Online Randomized Controlled Trial. J Am Med Dir Assoc 2015 Nov 1;16(11):990–997. [doi: 10.1016/j.jamda.2015.06.014]

21. DeRight J, Jorgensen RS. I Just Want My Research Credit: Frequency of Suboptimal Effort in a Non-Clinical Healthy Undergraduate Sample. Clin Neuropsychol 2014;0:1–17. [doi: 10.1080/13854046.2014.989267]

22. Healy AF, Kole JA, Buck-Gengler CJ, Bourne LE. Effects of Prolonged Work on Data Entry Speed and Accuracy. J Exp Psychol Appl 2004;10(3):188–199. [doi: 10.1037/1076-898X.10.3.188]

23. Rolstad S, Adler J, Rydén A. Response Burden and Questionnaire Length: Is Shorter Better? A Review and Meta-analysis. Value Health 2011 Dec;14(8):1101–1108. [doi: 10.1016/j.jval.2011.06.003]

24. McCambridge J, Kalaitzaki E, White IR, Khadjesari Z, Murray E, Linke S, Thompson SG, Godfrey C, Wallace P. Impact of Length or Relevance of Questionnaires on Attrition in Online Trials: Randomized Controlled Trial. J Med Internet Res 2011 Nov 18;13(4):e96. [doi: 10.2196/jmir.1733]

25. Donkin L, Christensen H, Naismith SL, Neal B, Hickie IB, Glozier N. A Systematic Review of the Impact of Adherence on the Effectiveness of e-Therapies. J Med Internet Res 2011 Aug 5;13(3). PMID:21821503

26. Spilgames. State of Online Gaming Report [Internet]. Spilgames; 2013. Available from: http://www.webcitation.org/6gt7YwEGt

27. Cowley B, Charles D, Black M, Hickey R. Toward an Understanding of Flow in Video Games. Comput Entertain 2008 Jul;6(2):20:1–20:27. [doi: 10.1145/1371216.1371223]

28. McGonigal J. Reality is broken: Why games make us better and how they can change the world. Penguin; 2011. ISBN:0-09-954028-2

29. Deterding S, Dixon D, Khaled R, Nacke L. From Game Design Elements to Gamefulness: Defining “Gamification.” Proc 15th Int Acad MindTrek Conf Envisioning Future Media Environ New York, NY, USA: ACM; 2011. p. 9–15. [doi: 10.1145/2181037.2181040]

30. Arai S, Sakamoto K, Washizaki H, Fukazawa Y. A Gamified Tool for Motivating Developers to Remove Warnings of Bug Pattern Tools. 2014 6th Int Workshop Empir Softw Eng Pract 2014. p. 37–42. [doi: 10.1109/IWESEP.2014.17]

31. Landers RN, Landers AK. An Empirical Test of the Theory of Gamified Learning: The Effect of Leaderboards on Time-on-Task and Academic Performance. Simul Gaming 2014 Dec 1;45(6):769–785. [doi: 10.1177/1046878114563662]

32. Ninaus M, Pereira G, Stefitz R, Prada R, Paiva A, Neuper C, Wood G. Game elements improve performance in a working memory training task. Int J Serious Games 2015 Feb 10;2(1). [doi: 10.17083/ijsg.v2i1.60]

33. Gustafsson A, Katzeff C, Bang M. Evaluation of a Pervasive Game for Domestic Energy Engagement Among Teenagers. Comput Entertain 2010 Jan;7(4):54:1–54:19. [doi: 10.1145/1658866.1658873]

34. Johnson D, Horton E, Mulcahy R, Foth M. Gamification and serious games within the domain of domestic energy consumption: A systematic review. Renew Sustain Energy Rev 2017 Jun 1;73:249–264. [doi: 10.1016/j.rser.2017.01.134]

35. Hamari J. Transforming homo economicus into homo ludens: A field experiment on gamification in a utilitarian peer-to-peer trading service. Electron Commer Res Appl 2013 Jul 1;12(4):236–245. [doi: 10.1016/j.elerap.2013.01.004]

36. Downes-Le Guin T, Baker R, Mechling J, Ruyle E. Myths and Realities of Respondent Engagement in Online Surveys. Int J Mark Res 2012 Sep 1;54(5):613–633. [doi: 10.2501/IJMR-54-5-613-633]

37. Sailer M, Hense JU, Mayr SK, Mandl H. How gamification motivates: An experimental study of the effects of specific game design elements on psychological need satisfaction. Comput Hum Behav 2017 Apr 1;69:371–380. [doi: 10.1016/j.chb.2016.12.033]

38. Seaborn K, Fels DI. Gamification in theory and action: A survey. Int J Hum-Comput Stud 2015 Feb 1;74(Supplement C):14–31. [doi: 10.1016/j.ijhcs.2014.09.006]

39. Hamari J, Koivisto J, Sarsa H. Does Gamification Work? – A Literature Review of Empirical Studies on Gamification. 2014 47th Hawaii Int Conf Syst Sci Los Alamitos, CA, USA: IEEE Computer Society; 2014. [doi: 10.1109/HICSS.2014.377]

40. Burke B. Gamification: Engagement strategies for business and IT. Gart Inc Novemb 2012;

41. LLC FM. JetBlue Badges gamification marketing fails to take off [Internet]. 2013 [cited 2018 Mar 13]. Available from: https://www.webinknow.com/2013/07/jetblue-badges-gamification-marketing-fails-to-take-off.html

42. Brown M, O'Neill N, van Woerden H, Eslambolchilar P, Jones M, John A. Gamification and Adherence to Web-Based Mental Health Interventions: A Systematic Review. JMIR Ment Health 2016 Aug 24;3(3). PMID:27558893

43. Looyestyn J, Kernot J, Boshoff K, Ryan J, Edney S, Maher C. Does gamification increase engagement with online programs? A systematic review. PloS One 2017;12(3):e0173403. PMID:28362821

44. Anguera JA, Boccanfuso J, Rintoul JL, Al-Hashimi O, Faraji F, Janowich J, Kong E, Larraburo Y, Rolle C, Johnston E, Gazzaley A. Video game training enhances cognitive control in older adults. Nature 2013 Sep 5;501(7465):97–101. [doi: 10.1038/nature12486]

45. Gamberini L, Cardullo S, Seraglia B, Bordin A. Neuropsychological testing through a Nintendo Wii console. Stud Health Technol Inform 2010;154:29–33.

46. Jaeggi SM, Buschkuehl M, Jonides J, Shah P. Short- and long-term benefits of cognitive training. Proc Natl Acad Sci 2011 Jun 21;108(25):10081–10086. [doi: 10.1073/pnas.1103228108]

47. Hawkins GE, Rae B, Nesbitt KV, Brown SD. Gamelike features might not improve data. Behav Res Methods 2013;45(2):301–318. [doi: 10.3758/s13428-012-0264-3]

48. McPherson J, Burns NR. Gs Invaders: Assessing a computer game-like test of processing speed. Behav Res Methods 2007;39(4):876–883. [doi: 10.3758/BF03192982]

49. Miranda AT, Palmer EM. Intrinsic motivation and attentional capture from gamelike features in a visual search task. Behav Res Methods 2014; 46: 159. [doi: 10.3758/s13428-013-0357-7]

50. Prins PJM, Dovis S, Ponsioen A, Brink E ten, van der Oord S. Does computerized working memory training with game elements enhance motivation and training efficacy in children with ADHD? Cyberpsychology Behav Soc Netw 2011 Mar;14(3):115–122. [doi: 10.1089/cyber.2009.0206]

51. Tong T, Chignell M. Developing a Serious Game for Cognitive Assessment: Choosing Settings and Measuring Performance. Proc Second Int Symp Chin CHI 2014. p. 70–79. [doi: 10.1145/2592235.2592246]

52. Dörrenbächer S, Müller PM, Tröger J, Kray J. Dissociable effects of game elements on motivation and cognition in a task-switching training in middle childhood. Cognition 2014;5:1275. [doi: 10.3389/fpsyg.2014.01275]

53. Washburn DA. The games psychologists play (and the data they provide). Behav Res Methods Instrum Comput 2003 May;35(2):185–193. [doi: 10.3758/BF03202541]

54. Borsboom D, Mellenbergh GJ. Test Validity in Cognitive Assessment. In: Leighton J, Gierl M, editors. Cogn Diagn Assess Educ Theory Appl Cambridge University Press; 2007. p. 85–116.

55. Lampit A, Hallock H, Valenzuela M. Computerized Cognitive Training in Cognitively Healthy Older Adults: A Systematic Review and Meta-Analysis of Effect Modifiers. PLOS Med 2014 Nov 18;11(11):e1001756. [doi: 10.1371/journal.pmed.1001756]

56. Jones A, Hardman CA, Lawrence N, Field M. Cognitive training as a potential treatment for overweight and obesity: A critical review of the evidence. Appetite 2018 May 1;124:50–67. PMID:28546010

57. Abikoff H. Cognitive Training Interventions in Children: Review of a New Approach. J Learn Disabil 1979 Feb 1;12(2):123–135. [doi: 10.1177/002221947901200213]

58. Melby-Lervåg M, Hulme C. Is working memory training effective? A meta-analytic review. Dev Psychol 2013 Feb;49(2):270–291. [doi: 10.1037/a0028228]

59. Amir B, Ralph P. Proposing a Theory of Gamification Effectiveness. Companion Proc 36th Int Conf Softw Eng New York, NY, USA: ACM; 2014. p. 626–627. [doi: 10.1145/2591062.2591148]

60. Marczewski A. A Revised Gamification Design Framework. Gamified UK - Gamification Expert [Internet] 2017 [cited 2018 Mar 14]; Available from: https://www.gamified.uk/2017/04/06/revised-gamification-design-framework/

61. Robson K, Plangger K, Kietzmann JH, McCarthy I, Pitt L. Is it all a game? Understanding the principles of gamification. Bus Horiz 2015 Jul 1;58(4):411–420. [doi: 10.1016/j.bushor.2015.03.006]

62. Aparicio AF, Vela FLG, Sánchez JLG, Montes JLI. Analysis and Application of Gamification. Proc 13th Int Conf InteracciÓN Pers-Ordenad [Internet] New York, NY, USA: ACM; 2012. p. 17:1–17:2. [doi: 10.1145/2379636.2379653]

63. Bartle R. Hearts, Clubs, Diamonds, Spades: Players Who Suit MUDs. J MUD Res 1996;

64. Hamari J, Tuunanen J. Player Types: A Meta-synthesis. Trans Digit Games Res Assoc 2014 Mar 24;1:29–53. [doi: 10.26503/todigra.v1i2.13]

65. Kahn AS, Shen C, Lu L, Ratan RA, Coary S, Hou J, Meng J, Osborn J, Williams D. The Trojan Player Typology: A cross-genre, cross-cultural, behaviorally validated scale of video game play motivations. Comput Hum Behav 2015 Aug;49:354–361. [doi: 10.1016/j.chb.2015.03.018]

66. Tondello GF, Wehbe RR, Diamond L, Busch M, Marczewski A, Nacke LE. The Gamification User Types Hexad Scale. Proc 2016 Annu Symp Comput-Hum Interact Play New York, NY, USA: ACM; 2016. p. 229–243. [doi: 10.1145/2967934.2968082]

67. Bavelier D, Green CS, Pouget A, Schrater P. Brain Plasticity Through the Life Span: Learning to Learn and Action Video Games. Annu Rev Neurosci 2012 Jun 20;35(1):391–416. [doi: 10.1146/annurev-neuro-060909-152832]

68. Przybylski AK, Wang JC. A large scale test of the gaming-enhancement hypothesis. PeerJ 2016 Nov 16;4:e2710. [doi: 10.7717/peerj.2710]

69. Lumsden J, Edwards EA, Lawrence NS, Coyle D, Munafò MR. Gamification of Cognitive Assessment and Cognitive Training: A Systematic Review of Applications and Efficacy. JMIR Serious Games 2016 Jul 15;4(2):e11. [doi: 10.2196/games.5888]

70. Dubbels B. Gamification, Serious Games, Ludic Simulation, and Other Contentious Categories. Int J Gaming Comput Mediat Simul 2013 Apr;5(2):1–19. [doi: 10.4018/jgcms.2013040101]

71. Sardi L, Idri A, Fernández-Alemán JL. A systematic review of gamification in e-health. J Biomed Inform 2017;71:31–48.

72. Boendermaker WJ, Boffo M, Wiers RW. Exploring Elements of Fun to Motivate Youth to Do Cognitive Bias Modification. Games Health J 2015 Sep 30;4(6):434–443. [doi: 10.1089/g4h.2015.0053]

73. Dunbar NE, Miller CH, Adame BJ, Elizondo J, Wilson SN, Lane BL, Kauffman AA, Bessarabova E, Jensen ML, Straub SK, Lee Y-H, Burgoon JK, Valacich JJ, Jenkins J, Zhang J. Implicit and explicit training in the mitigation of cognitive bias through the use of a serious game. Comput Hum Behav 2014;37:307–318. [doi: 10.1016/j.chb.2014.04.053]

74. Djaouti D, Alvarez J, Jessel J-P, Rampnoux O. Origins of serious games. Serious Games Edutainment Appl Springer; 2011. p. 25–43.

75. Donchin E. Video games as research tools: The Space Fortress game. Behav Res Methods Instrum Comput 1995 Jun;27(2):217–223. [doi: 10.3758/BF03204735]

76. Malone TW. What Makes Things Fun to Learn? Heuristics for Designing Instructional Computer Games. Proc 3rd ACM SIGSMALL Symp First SIGPC Symp Small Syst New York, NY, USA: ACM; 1980. p. 162–169. [doi: 10.1145/800088.802839]

77. Michael DR, Chen SL. Serious games: Games that educate, train, and inform. Muska & Lipman/Premier-Trade; 2005.

78. Connolly TM, Boyle EA, MacArthur E, Hainey T, Boyle JM. A systematic literature review of empirical evidence on computer games and serious games. Comput Educ 2012;59(2):661–686.

79. Werbach K. (Re)Defining Gamification: A Process Approach. Persuas Technol Springer, Cham; 2014. p. 266–272. [doi: 10.1007/978-3-319-07127-5_23]

80. Eriksson B, Musialik M, Wagner J. Gamification - Engaging the Future. 2012 Aug 6;

81. Liu Y, Alexandrova T, Nakajima T. Gamifying Intelligent Environments. Proc 2011 Int ACM Workshop Ubiquitous Meta User Interfaces New York, NY, USA: ACM; 2011. p. 7–12. [doi: 10.1145/2072652.2072655]

82. Massung E, Coyle D, Cater KF, Jay M, Preist C. Using Crowdsourcing to Support Pro-environmental Community Activism. Proc SIGCHI Conf Hum Factors Comput Syst New York, NY, USA: ACM; 2013. p. 371–380. [doi: 10.1145/2470654.2470708]

83. Bogost I. Gamification is bullshit. Gameful World Approaches Issues Appl 2015;65.

84. Deterding S, Björk SL, Nacke LE, Dixon D, Lawley E. Designing Gamification: Creating Gameful and Playful Experiences. CHI 13 Ext Abstr Hum Factors Comput Syst New York, NY, USA: ACM; 2013. p. 3263–3266. [doi: 10.1145/2468356.2479662]

85. Boulet G. Gamification: The Latest Buzzword and the Next Fad. eLearn 2012 Dec;2012(12). [doi: 10.1145/2407138.2421596]

86. Robertson M. Can't play, won't play. Hide Seek 2010;6:2010.

87. Cugelman B. Gamification: What It Is and Why It Matters to Digital Health Behavior Change Developers. JMIR Serious Games 2013 Dec 12;1(1):e3. [doi: 10.2196/games.3139]

88. Deterding S. Pawned. Gamification and Its Discontents [Internet]. 18:21:40 UTC. Available from: https://www.slideshare.net/dings/pawned-gamification-and-its-discontents/13-Ga_b_e_Z_ic

89. McPherson J, Burns NR. Assessing the validity of computer-game-like tests of processing speed and working memory. Behav Res Methods 2008 Nov;40(4):969–981. PMID:19001388

90. Trapp W, Hasmann A, Gallhofer B, Schwerdtner J, Guenther W, Dobmeier M. Cognitive remediation improves cognition and good cognitive performance increases time to relapse – results of a 5 year catamnestic study in schizophrenia patients. Clin Schizophr Relat Psychoses 2013;

91. Gamberini L, Martino F, Seraglia B, Spagnolli A, Fabregat M, Ibanez F, Alcaniz M, Andres JM. Eldergames project: An innovative mixed reality table-top solution to preserve cognitive functions in elderly people. 2nd Conf Hum Syst Interact 2009 HSI 09 2009. p. 164–169. [doi: 10.1109/HSI.2009.5090973]

92. Dovis, Oord, Wiers, Prins. Can Motivation Normalize Working Memory and Task Persistence in Children with Attention-Deficit/Hyperactivity Disorder? The Effects of Money and Computer-Gaming. J Abnorm Child Psychol 2011;40(5):669–681. [doi: 10.1007/s10802-011-9601-8]

93. Delisle J, Braun CMJ. A Context for Normalizing Impulsiveness at Work for Adults with Attention Deficit/Hyperactivity Disorder (Combined Type). Arch Clin Neuropsychol 2011;26:602–613.

94. Lim, Lee, Guan, Fung, Zhao, Teng, Zhang, Krishnan. A Brain-Computer Interface Based Attention Training Program for Treating Attention Deficit Hyperactivity Disorder. PLoS ONE 2012;

95. Heller MD, Roots K, Srivastava S, Schumann J, Srivastava J, Hale TS. A Machine Learning-Based Analysis of Game Data for Attention Deficit Hyperactivity Disorder Assessment. Games Health J 2013 Oct 1;2:291–298. [doi: 10.1089/g4h.2013.0058]

96. Verhaegh J, Fontijn WFJ, Aarts EHL, Resing WCM. In-game assessment and training of nonverbal cognitive skills using TagTiles. Pers Ubiquitous Comput 2013;17(8):1637–1646. [doi: 10.1007/s00779-012-0527-0]

97. Aalbers T, Baars MAE, Rikkert MGMO, Kessels RPC. Puzzling With Online Games (BAM-COG): Reliability, Validity, and Feasibility of an Online Self-Monitor for Cognitive Performance in Aging Adults. J Med Internet Res 2013 Dec;15(12). [doi: 10.2196/jmir.2860]

98. Fagundo, Santamaría, Forcano, Giner-Bartolomé, Jiménez-Murcia, Sánchez, Granero, Ben-Moussa, Magnenat-Thalmann, Konstantas, Lam, Lucas, Nielsen, Bults, Tarrega, Menchón, de la Torre, Cardi, Treasure, Fernández-Aranda. Video game therapy for emotional regulation and impulsivity control in a series of treated cases with bulimia nervosa. Eur Eat Disord Rev J Eat Disord Assoc 2013;21(6):493–499. [doi: 10.1002/erv.2259]

99. van der Oord S, Ponsioen AJGB, Geurts HM, Brink EL Ten, Prins PJM. A Pilot Study of the Efficacy of a Computerized Executive Functioning Remediation Training With Game Elements for Children With ADHD in an Outpatient Setting: Outcome on Parent- and TeacherRated Executive Functioning and ADHD Behavior. J Atten Disord 2014;18(8):699–712. [doi: 10.1177/1087054712453167]

100. Brown HR, Zeidman P, Smittenaar P, Adams RA, McNab F, Rutledge RB, Dolan RJ. Crowdsourcing for cognitive science--the utility of smartphones. PloS One 2014;9. [doi: 10.1371/journal.pone.0100662]

101. Katz B, Jaeggi S, Buschkuehl M, Stegman A, Shah P. Differential effect of motivational features on training improvements in school-based cognitive training. Front Hum Neurosci 2014;8:242. [doi: 10.3389/fnhum.2014.00242]

102. Lee T-S, Goh SJA, Quek SY, Phillips R, Guan C, Cheung YB, Feng L, Teng SSW, Wang CC, Chin ZY, Zhang H, Ng TP, Lee J, Keefe R, Krishnan KRR. A Brain-Computer Interface Based Cognitive Training System for Healthy Elderly: A Randomized Control Pilot Study for Usability and Preliminary Efficacy. PLoS ONE 2013 Nov 18;8(11):e79419. [doi: 10.1371/journal.pone.0079419]

103. Atkins SM, Sprenger AM, Colflesh GJH, Briner TL, Buchanan JB, Chavis SE, Chen S, Iannuzzi GL, Kashtelyan V, Dowling E, Harbison JI, Bolger DJ, Bunting MF, Dougherty MR. Measuring Working Memory Is All Fun and Games A Four-Dimensional Spatial Game Predicts Cognitive Task Performance. Exp Psychol 2014;61(6):417–438. [doi: 10.1027/1618-3169/a000262]

104. McNab F, Dolan RJ. Dissociating distractor-filtering at encoding and during maintenance. J Exp Psychol Hum Percept Perform 2014;40(3):960–967. [doi: 10.1037/a0036013]

105. O'Toole LJ, Dennis TA. Mental Health on the Go: Effects of a Gamified Attention-Bias Modification Mobile Application in Trait-Anxious Adults. Psychophysiology 2014;51:S71–S71.

106. Tenorio Delgado, Arango Uribe, Aparicio Alonso, Rosas Diaz. TENI: A comprehensive battery for cognitive assessment based on games and technology. Child Neuropsychol J Norm Abnorm Dev Child Adolesc 2014;1–16. [doi: 10.1080/09297049.2014.977241]

107. De Vries M, Prins PJM, Schmand BA, Geurts HM. Working memory and cognitive flexibility-training for children with an autism spectrum disorder: A randomized controlled trial. J Child Psychol Psychiatry 2015 May;56(5):566–576. [doi: 10.1111/jcpp.12324]

108. Dovis, Van Der Oord, Wiers, Prins. Improving executive functioning in children with ADHD: Training multiple executive functions within the context of a computer game. A

randomized double-blind placebo controlled trial. PLoS ONE 2015;10(4):e0121651. [doi: 10.1371/journal.pone.0121651]

109. Kim K-W, Choi Y, You H, Na DL, Yoh M-S, Park J-K, Seo J-H, Ko M-H. Effects of a Serious Game Training on Cognitive Functions in Older Adults. J Am Geriatr Soc 2015 Mar;63(3):603–605. [doi: 10.1111/jgs.13304]

110. Manera V, Petit P-D, Derreumaux A, Orvieto I, Romagnoli M, Lyttle G, David R, Robert PH. "Kitchen and cooking," a serious game for mild cognitive impairment and Alzheimer's disease: a pilot study. Front Aging Neurosci 2015 Mar 17;7. [doi: 10.3389/Fnagi.2015.00024]

111. Tarnanas I, Laskaris N, Tsolaki M, Muri R, Nef T, Mosimann UP. On the comparison of a novel serious game and electroencephalography biomarkers for early dementia screening. Adv Exp Med Biol 2015;821:63–77. [doi: 10.1007/978-3-319-08939-3_11]

112. Siraly E, Szabo A, Szita B, Kovacs V, Fodor Z, Marosi C, Salacz P, Hidasi Z, Maros V, Hanak P, Csibri E, Csukly G. Monitoring the Early Signs of Cognitive Decline in Elderly by Computer Games: An MRI Study. Plos One 2015 Feb 23;10(2):e0117918. [doi: 10.1371/journal.pone.0117918]

113. Colzato LS, van den Wildenberg WPM, Zmigrod S, Hommel B. Action video gaming and cognitive control: playing first person shooter games is associated with improvement in working memory but not action inhibition. Psychol Res 2013 Mar;77(2):234–239. PMID:22270615

114. Durlach PJ, Kring JP, Bowens LD. Effects of action video game experience on change detection. Mil Psychol 2009;21(1):24–39. [doi: 10.1080/08995600802565694]

115. Green CS, Bavelier D. Enumeration versus multiple object tracking: the case of action video game players. Cognition 2006;101(1):217–245. [doi: 10.1016/j.cognition.2005.10.004]

116. Dobrowolski P, Hanusz K, Sobczyk B, Skorko M, Wiatrow A. Cognitive enhancement in video game players: The role of video game genre. Comput Hum Behav 2015 Mar;44:59–63. [doi: 10.1016/j.chb.2014.11.051]

117. Dye MWG, Green CS, Bavelier D. Increasing Speed of Processing With Action Video Games. Curr Dir Psychol Sci 2009;18(6):321–326. PMID:20485453

118. Abikoff H. Efficacy of cognitive training interventions in hyperactive children: A critical review. Clinical Psychology Review, 5, 479-512. Clin Psychol Rev 1985;5:479–512. [doi: 10.1016/0272-7358(85)90005-4]

119. Burgess PW, Alderman N, Forbes C, CosTELlo A, M-A.coates L, Dawson DR, Anderson ND, Gilbert SJ, Dumontheil I, Channon S. The case for the development and use of measures of executive function in experimental and clinical neuropsychology. J Int Neuropsychol Soc 2006 Mar;12(02):194–209. [doi: 10.1017/S1355617706060310]

120. Kolb D. Experiential learning: Experience as the source of learning and development. Prentice Hall.; 1984.

121. Prins PJM, Brink ET, Dovis S, Ponsioen A, Geurts HM, de Vries M, van der Oord S. “Braingame Brian”: Toward an Executive Function Training Program with Game Elements for Children with ADHD and Cognitive Control Problems. Games Health J 2013;2(1):44–49. [doi: 10.1089/g4h.2013.0004]

122. Csikszentmihalyi M. Flow: The psychology of optimal experience. HarperPerennial New York; 1991.

123. Malone TW. Toward a theory of intrinsically motivating instruction. Cogn Sci 1981 Oct;5(4):333–369. [doi: 10.1016/S0364-0213(81)80017-1]

124. Deci EL, Ryan RM. Handbook of self-determination research. University Rochester Press; 2002. ISBN: 1580461565s

125. Kotler S. The Rise of Superman. 2014.

126. Jones MG. Creating Electronic Learning Environments: Games, Flow, and the User Interface. 1998.

127. Sweetser P, Wyeth P. GameFlow: A Model for Evaluating Player Enjoyment in Games. Comput Entertain 2005 Jul;3:3–3. [doi: 10.1145/1077246.1077253]

128. Smith BP. Flow and the enjoyment of video games. 2006.

129. Susan A. Jackson RCE. Assessing Flow in Physical Activity: The Flow State Scale-2 and Dispositional Flow Scale-2. Hum Kinet J. 2010.

130. Hamari J, Koivisto J. Measuring flow in gamification: Dispositional Flow Scale-2. Comput Hum Behav 2014;40:133–143. [doi: 10.1016/j.chb.2014.07.048]

131. Procci K, Singer AR, Levy KR, Bowers C. Measuring the flow experience of gamers: An evaluation of the DFS-2. Comput Hum Behav 2012;28(6):2306–2312. [doi: 10.1016/j.chb.2012.06.039]

132. Berlyne DE. Structure and direction in thinking. 1965;

133. Ryan RM, Deci EL. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. Am Psychol 2000;55(1):68–78. [doi: 10.1037/0003-066X.55.1.68]

134. Ryan RM, Rigby CS, Przybylski A. The Motivational Pull of Video Games: A Self-Determination Theory Approach. Motiv Emot 2006 Dec 1;30(4):344–360. [doi: 10.1007/s11031-006-9051-8]

135. Przybylski AK, Rigby CS, Ryan RM. A Motivational Model of Video Game Engagement. Rev Gen Psychol 2010;14(2):154–166.

136. Rigby S, Ryan RM. Glued to games: How video games draw us in and hold us spellbound. 2011;186.

137. Dunbar NE, Wilson SN, Adame BJ, Elizondo J, Jensen ML, Miller CH, Kauffman AA, Seltsam T, Bessarabova E, Vincent C, Straub SK, Ralston R, Dulawan CL, Ramirez D,

Squire K, Valacich JS, Burgoon JK. MACBETH: Development of a training game for the mitigation of cognitive bias. Int J Game-Based Learn 2013;3.

138. McAuley E, Duncan T, Tammen VV. Psychometric properties of the Intrinsic Motivation Inventory in a competitive sport setting: a confirmatory factor analysis. Res Q Exerc Sport 1989 Mar;60(1):48–58. PMID:2489825

139. Verhaegh J, Fontijn WFJ, Resing WCM. On the correlation between children's performances on electronic board tasks and nonverbal intelligence test measures. Comput Educ 2013 Nov;69:419–430. [doi: 10.1016/j.compedu.2013.07.026]

140. Kok AJ, Kong TY, Bernard-Opitz V. A Comparison of the Effects of Structured Play and Facilitated Play Approaches on Preschoolers with Autism A Case Study. Autism 2002 Jun 1;6(2):181–196. [doi: 10.1177/1362361302006002005]

141. Video Games and ADHD: What's the Link? [Internet]. WebMD. 2012 [cited 2015 Sep 22]. Available from: http://www.webcitation.org/6gt9IM416

142. Dougherty DD, Bonab AA, Spencer TJ, Rauch SL, Madras BK, Fischman AJ. Dopamine transporter density in patients with attention deficit hyperactivity disorder. The Lancet 1999 Dec 25;354(9196):2132–2133. [doi: 10.1016/S0140-6736(99)04030-1]

143. LaHoste, Swanson, Wigal, Glabe, Wigal, King, Kennedy. Dopamine D4 receptor gene polymorphism is associated with attention deficit hyperactivity disorder. Mol Psychiatry 1996 May;1(2):121–124.

144. Nieoullon A. Dopamine and the regulation of cognition and attention. Prog Neurobiol 2002 May;67(1):53–83. [doi: 10.1016/S0301-0082(02)00011-4]

145. Nieoullon A, Coquerel A. Dopamine: a key regulator to adapt action, emotion, motivation and cognition. Curr Opin Neurol 2003 Dec;16 Suppl 2:S3–9.

146. Axelson RD, Flick A. Defining Student Engagement. Change Mag High Learn 2010 Dec 27;43(1):38–43. [doi: 10.1080/00091383.2011.533096]

147. Macey William H., Schneider Benjamin. The Meaning of Employee Engagement. Ind Organ Psychol 2008 Feb 29;1(1):3–30. [doi: 10.1111/j.1754-9434.2007.0002.x]

148. Couper MP, Alexander GL, Zhang N, Little RJA, Maddy N, Nowak MA, McClure JB, Calvi JJ, Rolnick SJ, Stopponi MA, Cole Johnson C. Engagement and retention: measuring breadth and depth of participant use of an online intervention. J Med Internet Res 2010 Nov 18;12(4):e52. PMID:21087922

149. Boyle EA, Connolly TM, Hainey T, Boyle JM. Engagement in digital entertainment games: A systematic review. Comput Hum Behav 2012 May;28(3):771–780. [doi: 10.1016/j.chb.2011.11.020]

150. Perski O, Blandford A, West R, Michie S. Conceptualising engagement with digital behaviour change interventions: a systematic review using principles from critical interpretive synthesis. Transl Behav Med 2016 Dec 13;1–14. [doi: 10.1007/s13142-016-0453-1]

151. Niemiec CP, Ryan RM. Autonomy, competence, and relatedness in the classroom: Applying self-determination theory to educational practice. Sch Field 2009 Jul 1;7(2):133–144. [doi: 10.1177/1477878509104318]

152. Kato PM. What do you mean when you say your serious game has been validated? Experimental vs. Test Validity [Internet]. 2013. Available from: http://www.webcitation.org/6gt9POLIu

153. Kato PM. Evaluating Efficacy and Validating Games for Health. Games Health J 2012 Feb 1;1:74–76. [doi: 10.1089/g4h.2012.1017]

154. Mekler ED, Brühlmann F, Opwis K, Tuch AN. Do Points, Levels and Leaderboards Harm Intrinsic Motivation?: An Empirical Analysis of Common Gamification Elements. Proc First Int Conf Gameful Des Res Appl New York, NY, USA: ACM; 2013. p. 66–73. [doi: 10.1145/2583008.2583017]

155. Preist C, Massung E, Coyle D. Competing or Aiming to Be Average?: Normification As a Means of Engaging Digital Volunteers. Proc 17th ACM Conf Comput Support Coop Work Soc Comput New York, NY, USA: ACM; 2014. p. 1222–1233. [doi: 10.1145/2531602.2531615]

156. Lumsden J, Skinner A, Woods AT, Lawrence NS, Munafò M. The effects of gamelike features and test location on cognitive test performance and participant enjoyment. PeerJ 2016;4:e2184. [doi: 10.7717/peerj.2184]

157. Schreiner M, Reiss S, Schweizer K. Method Effects on Assessing Equivalence of Online and Offline Administration of a Cognitive Measure: The Exchange Test. Int J Internet Sci 2014;9(1):52–63.

158. Goodman JK, Cryder CE, Cheema A. Data Collection in a Flat World: The Strengths and Weaknesses of Mechanical Turk Samples. J Behav Decis Mak 2013 Jul 1;26(3):213–224. [doi: 10.1002/bdm.1753]

159. Lewis-Evans B. Gamasutra: Dopamine and games Liking, learning, or wanting to play? [Internet]. [cited 2018 Nov 10]. Available from: https://www.gamasutra.com/blogs/BenLewisEvans/20130827/198975/Dopamine_and_games__Liking_learning_or_wanting_to_play.php?print=1

160. Kiyatkin EA, Gratton A. Electrochemical monitoring of extracellular dopamine in nucleus accumbens of rats lever-pressing for food. Brain Res 1994 Aug 1;652(2):225–234. [doi: 10.1016/0006-8993(94)90231-3]

161. Hajnal A, Norgren R. Accumbens dopamine mechanisms in sucrose intake. Brain Res 2001 Jun 15;904(1):76–84. [doi: 10.1016/S0006-8993(01)02451-9]

162. Pickens R, Harris WC. Self-administration of d-amphetamine by rats. Psychopharmacologia 1968;12(2):158–163. PMID:5657050

163. Wise RA. Dopamine, learning and motivation. Nat Rev Neurosci 2004 Jun;5(6):483–494. [doi: 10.1038/nrn1406]

164. Koepp MJ, Gunn RN, Lawrence AD, Cunningham VJ, Dagher A, Jones T, Brooks DJ, Bench CJ, Grasby PM. Evidence for striatal dopamine release during a video game. Nature 1998 May 21;393(6682):266–268. PMID:9607763

165. Robinson S, Sandstrom SM, Denenberg VH, Palmiter RD. Distinguishing Whether Dopamine Regulates Liking, Wanting, and/or Learning About Rewards. Behav Neurosci 2005;119(1):5–15. [doi: 10.1037/0735-7044.119.1.5]

166. Berridge KC, Robinson TE. What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience? Brain Res Rev 1998 Dec 1;28(3):309–369. [doi: 10.1016/S0165-0173(98)00019-8]

167. Arias-Carrión O, Pöppel E. Dopamine, learning, and reward-seeking behavior. Acta Neurobiol Exp (Warsz) 2007;67(4):481–488. PMID:18320725

168. Arias-Carrión O, Stamelou M, Murillo-Rodríguez E, Menéndez-González M, Pöppel E. Dopaminergic reward system: a short integrative review. Int Arch Med 2010 Oct 6;3:24. PMID:20925949

169. Burgers C, Eden A, van Engelenburg MD, Buningh S. How feedback boosts motivation and play in a brain-training game. Comput Hum Behav 2015 Jul 1;48:94–103. [doi: 10.1016/j.chb.2015.01.038]

170. Azevedo R, Bernard RM. A Meta-Analysis of the Effects of Feedback in Computer-Based Instruction. J Educ Comput Res 1995 Sep 1;13(2):111–127. [doi: 10.2190/9LMD-3U28-3A0G-FTQT]

171. World Medical Association. WMA Declaration of Helsinki—ethical principles for medical research involving human subjects. 2013.

172. Knoeferle KM, Woods A, Käppler F, Spence C. That Sounds Sweet: Using Cross-Modal Correspondences to Communicate Gustatory Attributes. Psychol Mark 2015 Jan 1;32(1):107–120. [doi: 10.1002/mar.20766]

173. Michel C, Woods AT, Neuhäuser M, Landgraf A, Spence C. Rotating plates: Online study demonstrates the importance of orientation in the plating of food. Food Qual Prefer 2015 Sep;44:194–202. [doi: 10.1016/j.foodqual.2015.04.015]

174. Verbruggen F, Logan GD. Automatic and controlled response inhibition: associative learning in the go/no-go and stop-signal paradigms. 2008; [doi: 10.1037/a0013170]

175. Benikos N, Johnstone SJ, Roodenrys SJ. Varying task difficulty in the Go/Nogo task: The effects of inhibitory control, arousal, and perceived effort on ERP components. Int J Psychophysiol 2013 Mar;87(3):262–272. [doi: 10.1016/j.ijpsycho.2012.08.005]

176. Bowley C, Faricy C, Hegarty B, J. Johnstone S, L. Smith J, J. Kelly P, A. Rushby J. The effects of inhibitory control training on alcohol consumption, implicit alcohol-related cognitions and brain electrical activity. Int J Psychophysiol 2013 Sep;89(3):342–348. [doi: 10.1016/j.ijpsycho.2013.04.011]

177. Kertzman S, Lowengrub K, Aizer A, Vainder M, Kotler M, Dannon PN. Go–no-go performance in pathological gamblers. Psychiatry Res 2008 Oct 30;161(1):1–10. [doi: 10.1016/j.psychres.2007.06.026]

178. Watson TD, Garvey KT. Neurocognitive correlates of processing food-related stimuli in a Go/No-go paradigm. Appetite 2013 Dec 1;71:40–47. [doi: 10.1016/j.appet.2013.07.007]

179. Yechiam E, Goodnight J, Bates JE, Busemeyer JR, Dodge KA, Pettit GS, Newman JP. A Formal Cognitive Model of the Go/No-Go Discrimination Task: Evaluation and Implications. Psychol Assess 2006 Sep;18(3):239–249.

180. Moller AC, Elliot AJ, Maier MA. Basic hue-meaning associations. Emot Wash DC 2009 Dec;9(6):898–902. PMID:20001133

181. Guitart-Masip M, Huys QJM, Fuentemilla L, Dayan P, Duzel E, Dolan RJ. Go and no-go learning in reward and punishment: interactions between affect and effect. NeuroImage 2012 Aug 1;62(1):154–166. [doi: 10.1016/j.neuroimage.2012.04.024]

182. Berger JO, Sellke T. Testing a Point Null Hypothesis: The Irreconcilability of P Values and Evidence. J Am Stat Assoc 1987 Mar 1;82(397):112–122. [doi: 10.1080/01621459.1987.10478397]

183. Blackwelder WC. “Proving the null hypothesis” in clinical trials. Control Clin Trials 1982 Dec;3(4):345–353. [doi: 10.1016/0197-2456(82)90024-1]

184. Rouder JN, Speckman PL, Sun D, Morey RD, Iverson G. Bayesian t tests for accepting and rejecting the null hypothesis. Psychon Bull Rev 2009 Apr;16(2):225–237. [doi: 10.3758/PBR.16.2.225]

185. Wetzels R, Raaijmakers JGW, Jakab E, Wagenmakers E-J. How to quantify support for and against the null hypothesis: A flexible WinBUGS implementation of a default Bayesian t test. Psychon Bull Rev 2009 Aug;16(4):752–760. [doi: 10.3758/PBR.16.4.752]

186. Raftery AE. Bayesian Model Selection in Social Research. Sociol Methodol 1995;25:111–163. [doi: 10.2307/271063]

187. Jeffreys H. Theory of Probability. 3rd ed. Oxford: Clarendon Press; 1961.

188. Hauser DJ, Schwarz N. Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. Behav Res Methods 2015 Mar 12;48(1):400–407. [doi: 10.3758/s13428-015-0578-z]

189. Wickelgren WA. Speed-accuracy tradeoff and information processing dynamics. Acta Psychol (Amst) 1977 Feb;41(1):67–85. [doi: 10.1016/0001-6918(77)90012-9]

190. Bogacz R. Speed-Accuracy Tradeoff. In: Jaeger D, Jung R, editors. Encycl Comput Neurosci New York, NY: Springer New York; 2015. p. 2798–2801.

191. Neath I, Earle A, Hallett D, Surprenant AM. Response time accuracy in Apple Macintosh computers. Behav Res Methods 2011 Mar 17;43(2):353–362. [doi: 10.3758/s13428-011-0069-9]

192. Plant RR, Turner G. Millisecond precision psychological research in a world of commodity computers: New hardware, new problems? Behav Res Methods 2009 Aug;41(3):598–614. [doi: 10.3758/BRM.41.3.598]

193. Stewart N, Chandler J, Paolacci G. Crowdsourcing Samples in Cognitive Science. Trends Cogn Sci 2017 Oct 1;21(10):736–748. [doi: 10.1016/j.tics.2017.06.007]

194. Reimers S, Stewart N. Presentation and response timing accuracy in Adobe Flash and HTML5/JavaScript Web experiments. Behav Res Methods 2015 Jun;47(2):309–327. PMID:24903687

195. de Leeuw JR, Motz BA. Psychophysics in a Web browser? Comparing response times collected with JavaScript and Psychophysics Toolbox in a visual search task. Behav Res Methods 2016 Mar;48(1):1–12. PMID:25761390

196. McDermott JM, Pérez-Edgar K, Fox NA. Variations of the flanker paradigm: Assessing selective attention in young children. Behav Res Methods 2007 Feb;39(1):62–70. [doi: 10.3758/BF03192844]

197. Prins, Ponsioen, van der Oord, Dovis. Does computerized working memory training with game elements enhance motivation and training efficacy in children with ADHD? Cyberpsychology Behav Soc Netw 2011 Mar;

198. Boendermaker WJ, Prins PJM, Wiers RW. Cognitive Bias Modification for adolescents with substance use problems – Can serious games help? J Behav Ther Exp Psychiatry 2015 Dec;49, Part A:13–20. [doi: 10.1016/j.jbtep.2015.03.008]

199. Peters SE, Lumsden J, Peh OH, Penton-Voak IS, Munafò MR, Robinson OJ. Cognitive bias modification for facial interpretation: a randomized controlled trial of transfer to self-report and cognitive measures in a healthy sample. Open Sci 2017 Dec 1;4(12):170681. [doi: 10.1098/rsos.170681]

200. Looi CY, Lumsden J, Suddell S, Granger K, Barnett JH, Munafo M, Voak IP. Real-world trial to examine the effectiveness and transference of cognitive bias modification to functional outcomes: study protocol for an RCT in volunteers taking anti-depressants. Open Sci Framew [Internet] 2017 Nov 7; [doi: None]

201. Looi CY, Lumsden J, Müller-Glodde M, Robinson OJ, Munafo M, Voak IP. Real-world trial to examine the effectiveness and transference of cognitive bias modification to cognitive and self-report measures: study protocol for a randomized control trial in healthy volunteers. Open Sci Framew [Internet] 2017 Jun 27; [doi: None]

202. Stewart N, Ungemach C, Harris AJL, Bartels DM, Newell BR, Paolacci G, Chandler J. The average laboratory samples a population of 7,300 Amazon Mechanical Turk workers. Judgm Decis Mak 2015 Sep;10(5):479–491.

203. Huff C, Tingley D. “Who are these people?” Evaluating the demographic characteristics and political preferences of MTurk survey respondents. Res Polit 2015 Aug 20;2(3):2053168015604648. [doi: 10.1177/2053168015604648]

204. Horton JJ, Chilton LB. The Labor Economics of Paid Crowdsourcing. Proc 11th ACM Conf Electron Commer [Internet] New York, NY, USA: ACM; 2010. p. 209–218. [doi: 10.1145/1807342.1807376]

205. Palan S, Schitter C. Prolific.ac—A subject pool for online experiments. J Behav Exp Finance 2018 Mar 1;17:22–27. [doi: 10.1016/j.jbef.2017.12.004]

206. Casler K, Bickel L, Hackett E. Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. Comput Hum Behav 2013 Nov 1;29(6):2156–2160. [doi: 10.1016/j.chb.2013.05.009]

207. Chandler J, Mueller P, Paolacci G. Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. Behav Res Methods 2014 Mar 1;46(1):112–130. [doi: 10.3758/s13428-013-0365-7]

208. The Internet's hidden science factory [Internet]. PBS NewsHour. 2015 [cited 2018 May 29]. Available from: https://www.pbs.org/newshour/science/inside-amazons-hidden-science-factory

209. Data Protection Act. c 29 1998.

210. Lumsden J, Skinner A, Coyle D, Lawrence N, Munafo M. Attrition from Web-Based Cognitive Testing: A Repeated Measures Comparison of Gamification Techniques. J Med Internet Res 2017;19(11):e395. [doi: 10.2196/jmir.8473]

211. Lodwick RK. Crossover designs: issues in construction, use, and communication [Thesis]. Queen Mary University of London; 2016.

212. Mekler ED, Brühlmann F, Tuch AN, Opwis K. Towards understanding the effects of individual gamification elements on intrinsic motivation and performance. Comput Hum Behav 2017 Jun 1;71:525–534. [doi: 10.1016/j.chb.2015.08.048]

213. Logan GD. On the ability to inhibit thought and action: A users' guide to the stop signal paradigm. In: Dagenbach D, Carr TH, editors. Inhib Process Atten Mem Lang San Diego, CA, US: Academic Press; 1994. p. 189–239.

214. Logan GD, Cowan WB. On the ability to inhibit thought and action: A theory of an act of control. Psychol Rev 1984;91(3):295–327. [doi: 10.1037/0033-295X.91.3.295]

215. Verbruggen F, Logan GD. Response inhibition in the stop-signal paradigm. Trends Cogn Sci 2008;12(11):418–424. [doi: 10.1016/j.tics.2008.07.005]

216. Schachar R, Logan GD, Robaey P, Chen S, Ickowicz A, Barr C. Restraint and Cancellation: Multiple Inhibition Deficits in Attention Deficit Hyperactivity Disorder. J Abnorm Child Psychol 2007;35(2):229–238. [doi: 10.1007/s10802-006-9075-2]

217. Cantab Research Suite [Internet]. cambridgecognition.com. 2014 [cited 2014 Oct 20]. Available from: http://www.cambridgecognition.com/academic/products

218. CANTAB Stop Signal Task [Internet]. 2017 [cited 2017 Sep 13]. Available from: http://www.webcitation.org/6tRv9eeZj

219. Logan GD, Schachar RJ, Tannock R. Impulsivity and Inhibitory Control. Psychol Sci 1997;8(1):60–64. [doi: 10.1111/j.1467-9280.1997.tb00545.x]

220. Band GPH, van der Molen MW, Logan GD. Horse-race model simulations of the stop-signal procedure. Acta Psychol (Amst) 2003;112(2):105–142. [doi: 10.1016/S0001-6918(02)00079-3]

221. Schell J. The Art of Game Design. CRC Press; 2008. ISBN:978-0-12-369496-6

222. Andreou P, Neale BM, Chen W, Christiansen H, Gabriels I, Heise A, Meidad S, Muller UC, Uebel H, Banaschewski T, Manor I, Oades R, Roeyers H, Rothenberger A, Sham P, Steinhausen H-C, Asherson P, Kuntsi J. Reaction time performance in ADHD: improvement under fast-incentive condition and familial effects. Psychol Med 2007 Dec;37(12):1703–1715. [doi: 10.1017/S0033291707000815]

223. Garrett DD, MacDonald SWS, Craik FIM. Intraindividual reaction time variability is malleable: feedback- and education-related reductions in variability with age. Front Hum Neurosci 2012;6:101. [doi: 10.3389/fnhum.2012.00101]

224. Verbruggen F, Logan GD. Models of Response Inhibition in the Stop-Signal and Stop-Change Paradigms. Neurosci Biobehav Rev 2009;33(5):647–661. [doi: 10.1016/j.neubiorev.2008.08.014]

225. Boendermaker WJ, Maceiras SS, Boffo M, Wiers RW. Attentional Bias Modification With Serious Game Elements: Evaluating the Shots Game. JMIR Serious Games 2016;4(2):e20. [doi: 10.2196/games.6464]

226. Deci EL, Ryan RM, Koestner R. A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. Psychol Bull 1999;125(6):627–668.

227. Deterding S. Situated motivational affordances of game elements: A conceptual model. Gamification Using Game Des Elem Non-Gaming Contexts Workshop CHI 2011.

228. Khadjesari Z, Murray E, Kalaitzaki E, White IR, McCambridge J, Thompson SG, Wallace P, Godfrey C. Impact and Costs of Incentives to Reduce Attrition in Online Trials: Two Randomized Controlled Trials. J Med Internet Res 2011 Mar 2;13(1):e26. [doi: 10.2196/jmir.1523]

229. Lee D, LaRose R. A Socio-Cognitive Model of Video Game Usage. J Broadcast Electron Media 2007;51:632–650.

230. Attali Y, Arieli-Attali M. Gamification in assessment: Do points affect test performance? Comput Educ 2015 Apr;83:57–63. [doi: 10.1016/j.compedu.2014.12.012]

231. Leotti LA, Wager TD. Motivational influences on response inhibition measures. J Exp Psychol Hum Percept Perform 2010 Apr;36(2):430–447. [doi: 10.1037/a0016802]

232. Boendermaker WJ, Gladwin TE, Peeters M, Prins PJM, Wiers RW. Training Working Memory in Adolescents Using Serious Game Elements: Pilot Randomized Controlled Trial. JMIR Serious Games 2018;6(2):e10. [doi: 10.2196/games.8364]

233. Schønau-Fog H, Bjørner T. “Sure, I Would Like to Continue” A Method for Mapping the Experience of Engagement in Video Games. Bull Sci Technol Soc 2012 Oct 1;32(5):405–412. [doi: 10.1177/0270467612469068]

234. Ariely D, Kamenica E, Prelec D. Man’s search for meaning: The case of Legos. J Econ Behav Organ 2008 Sep 1;67(3):671–677. [doi: 10.1016/j.jebo.2008.01.004]

235. Ryan RM, Deci EL. Intrinsic Motivation Inventory (IMI). 1995.

236. de Boer MR, Waterlander WE, Kuijper LD, Steenhuis IH, Twisk JW. Testing for baseline differences in randomized controlled trials: an unhealthy research behavior that is hard

to eradicate. Int J Behav Nutr Phys Act 2015 Jan 24;12:4. [doi: 10.1186/s12966-015-0162-z]

237. Kaufmann N, Schulze T, Veit D. More than fun and money. Worker Motivation in Crowdsourcing--A Study on Mechanical Turk. Proc Seventeenth Am Conf Inf Syst 2011.

238. Ye T, You S, Robert Jr L. When Does More Money Work? Examining the Role of Perceived Fairness in Pay on the Performance Quality of Crowdworkers. AAAI; 2017.

239. Lovett M, Bajaba S, Lovett M, Simmering MJ. Data Quality from Crowdsourced Surveys: A Mixed Method Inquiry into Perceptions of Amazon's Mechanical Turk Masters. Appl Psychol 2018 Apr 1;67(2):339–366. [doi: 10.1111/apps.12124]

240. Silberman MS, Tomlinson B, LaPlante R, Ross J, Irani L, Zaldivar A. Responsible Research with Crowds: Pay Crowdworkers at Least Minimum Wage. Commun ACM 2018 Feb;61(3):39–41. [doi: 10.1145/3180492]

241. Gleibs IH. Collecting data using crowdsourcing marketplaces raises ethical questions for academic researchers [Internet]. Impact Soc Sci Blog. 2016 [cited 2018 May 20]. Available from: http://blogs.lse.ac.uk/impactofsocialsciences

242. Zhang P. Motivational Affordances: Fundamental Reasons for ICT Design and Use [Internet]. Rochester, NY: Social Science Research Network; 2008. Report No.: ID 2352593. Available from: https://papers.ssrn.com/abstract=2352593

243. Peng W, Lin J-H, Pfeiffer KA, Winn B. Need Satisfaction Supportive Game Features as Motivational Determinants: An Experimental Study of a Self-Determination Theory Guided Exergame. Media Psychol 2012 May 18;15(2):175–196. [doi: 10.1080/15213269.2012.673850]

244. Baranowski MT, Lu AS, Buday R, Lyons EJ, Schell J, Russoniello C. Stories in Games for Health: More Pros or Cons? Games Health J 2013 Oct;2(5):256–263. PMID:26196925

245. Nacke L, Deterding S. The maturing of gamification research. Comput Hum Behav 2017 Jan 11;71. [doi: 10.1016/j.chb.2016.11.062]

246. Johnson D, Deterding S, Kuhn K-A, Staneva A, Stoyanov S, Hides L. Gamification for health and wellbeing: A systematic review of the literature. Internet Interv 2016 Nov 1;6:89–106. [doi: 10.1016/j.invent.2016.10.002]

247. Sebastian Deterding. Gamification for Health Behaviour Change [Internet]. 13:18:51 UTC. Available from: https://www.slideshare.net/dings/gamification-for-health-behaviour-change-88500762

248. Roulin N. Don't Throw the Baby Out With the Bathwater: Comparing Data Quality of Crowdsourcing, Online Panels, and Student Samples. Ind Organ Psychol 2015 Jun;8(2):190–196. [doi: 10.1017/iop.2015.24]

249. Clifford S, Jewell RM, Waggoner PD. Are samples drawn from Mechanical Turk valid for research on political ideology? Res Polit 2015 Oct 1;2(4):2053168015622072. [doi: 10.1177/2053168015622072]

250. Rogstadius J, Kostakos V, Kittur A, Smus B, Laredo J, Vukovic M. An Assessment of Intrinsic and Extrinsic Motivation on Task Performance in Crowdsourcing Markets. Proc ICWSM11 2011. [doi: 10.13140/RG.2.2.19170.94401]

251. Jiang L, Wagner C, Nardi B. Not Just in it for the Money: A Qualitative Investigation of Workers' Perceived Benefits of Micro-task Crowdsourcing. 2015 48th Hawaii Int Conf Syst Sci 2015. p. 773–782. [doi: 10.1109/HICSS.2015.98]

252. Morschheuser B, Hamari J, Koivisto J. Gamification in Crowdsourcing: A Review. Proc 2016 49th Hawaii Int Conf Syst Sci HICSS [Internet] Washington, DC, USA: IEEE Computer Society; 2016. p. 4375–4384. [doi: 10.1109/HICSS.2016.543]

253. Chandler D, Kapelner A. Breaking Monotony with Meaning: Motivation in Crowdsourcing Markets. J Econ Behav Organ 2013 Jun;90:123–133. [doi: 10.1016/j.jebo.2013.03.003]

254. Goswami I, Urminsky O. Figuring Out Preference or Balancing Out Effort: Do Inferences From Incentives Undermine Post-Incentive Motivation? 2018;

255. Short-term rewards don't sap long-term motivation [Internet]. Chic Booth Rev. [cited 2018 May 17]. Available from: http://review.chicagobooth.edu/behavioral-science/2018/article/short-term-rewards-don-t-sap-long-term-motivation

256. Brewer R, Anthony L, Brown Q, Irwin G, Nias J, Tate B. Using Gamification to Motivate Children to Complete Empirical Studies in Lab Environments. Proc 12th Int Conf Interact Des Child New York, NY, USA: ACM; 2013. p. 388–391. [doi: 10.1145/2485760.2485816]

257. Denny P. The Effect of Virtual Achievements on Student Engagement. Proc SIGCHI Conf Hum Factors Comput Syst New York, NY, USA: ACM; 2013. p. 763–772. [doi: 10.1145/2470654.2470763]

258. Barata G, Gama S, Jorge J, Goncalves D. Engaging Engineering Students with Gamification. 2013 5th Int Conf Games Virtual Worlds Serious Appl VS-GAMES 2013. p. 1–8. [doi: 10.1109/VS-GAMES.2013.6624228]

259. A User-Centered Theoretical Framework for Meaningful Gamification [Internet]. [cited 2015 Sep 15]. Available from: http://www.bgnlab.ca/blog/2012/6/13/a-user-centered-theoretical-framework-for-meaningful-gamific.html

260. Rigby S, Ryan R. The player experience of need satisfaction (PENS) model. Immersyve Inc 2007;1–22.

261. Blohm I, Leimeister JM. Gamification. Bus Inf Syst Eng 2013 Aug 1;5(4):275–278. [doi: 10.1007/s12599-013-0273-5]

262. Sakamoto M, Nakajima T, Alexandrova T. Value-Based Design for Gamifying Daily Activities. Entertain Comput - ICEC 2012 Springer, Berlin, Heidelberg; 2012. p. 421–424. [doi: 10.1007/978-3-642-33542-6_43]

263. Mora A, Riera D, Gonzalez C, Arnedo-Moreno J. A Literature Review of Gamification Design Frameworks. 2015 7th Int Conf Games Virtual Worlds Serious Appl VS-Games 2015. p. 1–8. [doi: 10.1109/VS-GAMES.2015.7295760]

264. Haaranen L, Ihantola P, Hakulinen L, Korhonen A. How (Not) to Introduce Badges to Online Exercises. Proc 45th ACM Tech Symp Comput Sci Educ New York, NY, USA: ACM; 2014. p. 33–38. [doi: 10.1145/2538862.2538921]

265. Michie S, Richardson M, Johnston M, Abraham C, Francis J, Hardeman W, Eccles MP, Cane J, Wood CE. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med Publ Soc Behav Med 2013 Aug;46(1):81–95. PMID:23512568

266. Linstone HA, Turoff M, others. The delphi method. Addison-Wesley Reading, MA; 1975.

267. Brockmyer JH, Fox CM, Curtiss KA, McBroom E, Burkhart KM, Pidruzny JN. The development of the Game Engagement Questionnaire: A measure of engagement in video game-playing. J Exp Soc Psychol 2009 Jul 1;45(4):624–634. [doi: 10.1016/j.jesp.2009.02.016]

268. Lafreniere M-AK, Verner-Filion J, Vallerand RJ. Development and validation of the Gaming Motivation Scale (GAMS). Individ Differ 2012 Jan 1;53:827–831.

269. Wu J-H, Wang S-C. What drives mobile commerce?: An empirical evaluation of the revised technology acceptance model. Inf Manage 2005 Jul 1;42(5):719–729. [doi: 10.1016/j.im.2004.07.001]

270. Wiebe EN, Lamb A, Hardy M, Sharek D. Measuring engagement in video game-based environments: Investigation of the User Engagement Scale. Comput Hum Behav 2014 Mar 1;32:123–132. [doi: 10.1016/j.chb.2013.12.001]

271. Jennett C, Cox AL, Cairns P, Dhoparee S, Epps A, Tijs T, Walton A. Measuring and defining the experience of immersion in games. Int J Hum-Comput Stud 2008 Sep;66(9):641–661. [doi: 10.1016/j.ijhcs.2008.04.004]

272. Boyle EA, Hainey T, Connolly TM, Gray G, Earp J, Ott M, Lim T, Ninaus M, Ribeiro C, Pereira J. An update to the systematic literature review of empirical evidence of the impacts and outcomes of computer games and serious games. Comput Educ 2016 Mar;94:178–192. [doi: 10.1016/j.compedu.2015.11.003]

273. Danaher BG, Boles SM, Akers L, Gordon JS, Severson HH. Defining Participant Exposure Measures in Web-Based Health Behavior Change Programs. J Med Internet Res 2006 Aug 30;8(3):e15. [doi: 10.2196/jmir.8.3.e15]

274. Lang A. The Limited Capacity Model of Mediated Message Processing. J Commun 2000 Mar 1;50(1):46–70. [doi: 10.1111/j.1460-2466.2000.tb02833.x]

275. Bezdek MA, Gerrig RJ. When Narrative Transportation Narrows Attention: Changes in Attentional Focus During Suspenseful Film Viewing. Media Psychol 2017 Jan 2;20(1):60–89. [doi: 10.1080/15213269.2015.1121830]

276. Cox AL, Cairns P, Berthouze N, Jennett C. The use of eyetracking for measuring immersion.

277. Lieberoth A. Shallow Gamification: Testing Psychological Effects of Framing an Activity as a Game. Games Cult 2015 May 1;10(3):229–248. [doi: 10.1177/1555412014559978]

278. Sun E, Jones B, Traca S, Bos MW. Leaderboard Position Psychology: Counterfactual Thinking. Proc 33rd Annu ACM Conf Ext Abstr Hum Factors Comput Syst [Internet] New York, NY, USA: ACM; 2015 [cited 2016 Mar 4]. p. 1217–1222. [doi: 10.1145/2702613.2732732]

279. Shen L, Hsee CK. Numerical Nudging: Using an Accelerating Score to Enhance Performance. Psychol Sci 2017 Aug 1;28(8):1077–1086. [doi: 10.1177/0956797617700497]

280. Diedenhofen B, Musch J. cocron: A Web Interface and R Package for the Statistical Comparison of Cronbach's Alpha Coefficients. Int J Internet Sci 2016;11(1).

281. Whiteside SP, Lynam DR. The Five Factor Model and impulsivity: using a structural model of personality to understand impulsivity. Personal Individ Differ 2001 Mar;30(4):669–689. [doi: 10.1016/S0191-8869(00)00064-7]

Appendices

Appendix A

Reasons for using gamification in cognitive training and testing

Reason	Game
To increase suitability for target age	Eldergames, Smart Harmony, Groundskeeper, Whack-a-mole, Wii Tests, Tap the Hedgehog, TENI
To reduce participant drop-out	Xcog, Cogoland, BAM-COG, ABMT App, The Great Brain Experiment, WMTrainer, Card-Pairing
To Increase suitability for target disorder	Megabot, Supermecha, Braingame Brian, Retirement Party,
To increase ecological-validity	Playmancer, Kitchen and Cooking, VAP-M, Neuroracer, MACBETH,
To increase short term engagement	Shapebuilder, Space Code, Space Matrix
To investigate the effects of game-like tasks	Ghost Trap, EM-Ants, Visual Search, Watermons

Appendix B

Games categorised by the age group they were aimed at

Age Group Targeted	Game	Count
All ages	EM-ANTS, Ghost-Trap, MACBETH, Playmancer, Shapebuilder, Space Code, Space Matrix, The Great Brain Experiment, Visual Search, Xcog	10
Children	Braingame Brian, WMTrainer, Groundskeeper, Megabot, Supermecha, Tap, TENI, Watermons, Cogoland	9
Adults	ABMTApp, Retirement Party	2
The Elderly	BAM-COG, Eldergames, Neuroracer, Wii Tests, Kitchen and Cooking, VAP-M Whack-a-mole, Smart Harmony, Card-Pairing	9

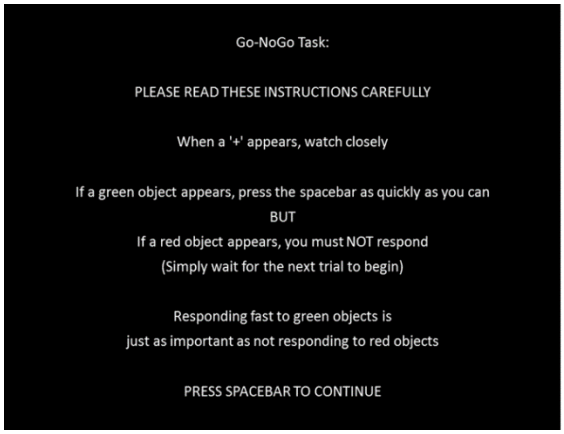
Appendix C

Games listed by category: testing or training

Category	Game	Count
Training	Supermecha, Megabot, Braingame Brian, Cogoland, Xcog, Smart Harmony, ABMTApp, WMTrainer, MACBETH, Playmancer, Neuroracer, Card-Pairing, Watermons	13
Testing	Space Code, Space Matrix, Eldergames, Wii Tests, Retirement Party, Groundskeeper, EM-Ants, Tap the Hedgehog, BAM-COG, VAP-M, The Great Brain Experiment, Whack-a-mole, Visual Search, Shapebuilder, TENI, Ghost Trap	16
Both	Kitchen and Cooking	1

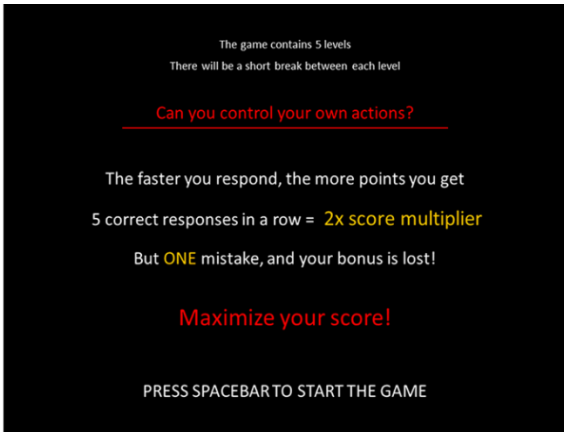
Appendix D

Instructions for the non-game variant in the GNG task



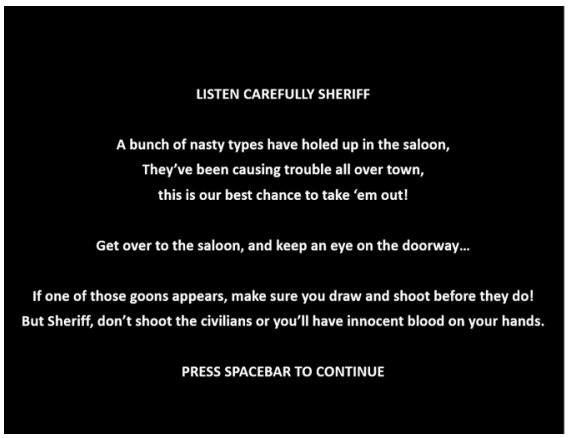
Appendix E

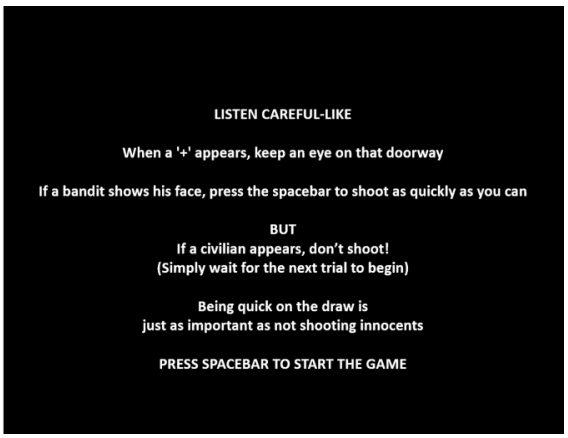
Instructions for the points variant in the GNG task



Appendix F

Instructions for the theme variant in the GNG task



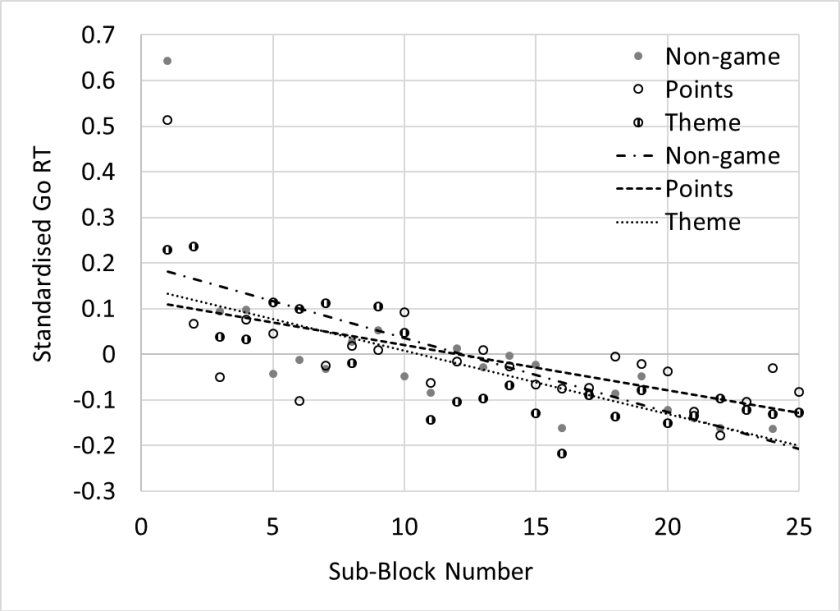


Appendix G

Experiment 1: Do reaction times get longer as the test goes on? Does this effect differ between task variants?

I calculated standardised Go RTs for based on the participants' mean RT and standard deviation of RT. I calculated mean standardised RTs for each sub-block combined over sites (12 trials per sub-block, 300 trials total).

A two-way ANOVA of standardised Go RT indicated strong evidence of a large effect of sub-block number ($F_{24,75}$=14.547, p<.001, η^2=.82), but not of task variant or an interaction (ps>.057). I used linear regression to assess the relationship between sub-block number and standardised Go RT in each task variant, and found strong evidence for a medium association in each case (non-game: r^2=.44, $F_{1,48}$=37.742, p<.001; points: r^2=.28, $F_{1,48}$=18.752, p<.001; theme: r^2=.58, $F_{1,48}$=65.578, p<.001). I scatterplotted the data and regression lines (Appendix G Figure 1).



Appendix G, Figure 1: *Standardised Go RT plotted against sub-block number, showing a downward trend as the experiment progresses. Shown separately by task variant.*

Appendix H

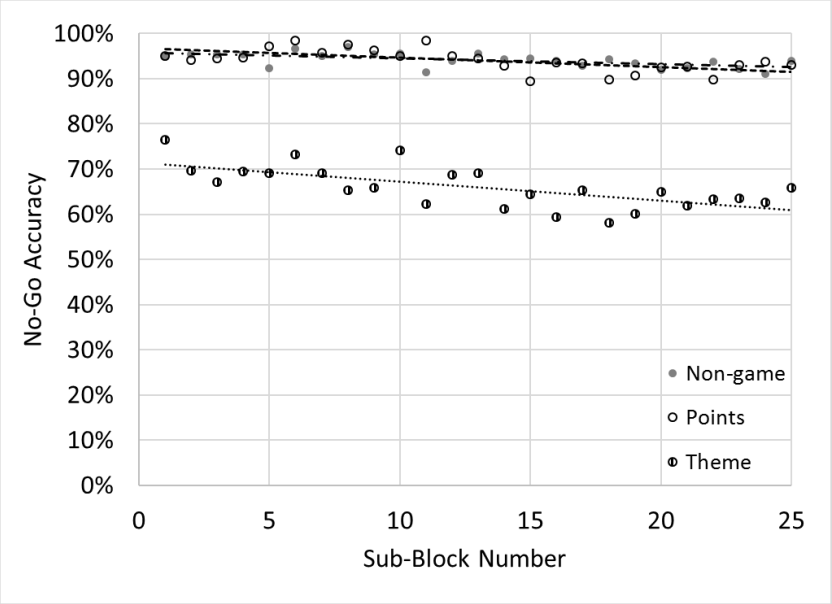
Results of Mann-Whitney u tests to confirm ANOVA findings on Go-trial accuracy

Variable	**Task Variant (N)**	**Task Variant (N)**	**U**	**Z-score**	***p***	***r***
Go Accuracy	Theme (93)	Points (99)	648.0	10.27	<.001	.74
Go Accuracy	Theme (93)	Non-game (95)	892.5	9.45	<.001	.69
Go Accuracy	Points (99)	Non-game (95)	4277.5	1.09	.275	.08
No-Go Accuracy	Theme (93)	Points (99)	169.5	11.52	<.001	.83
No-Go Accuracy	Theme (93)	Non-game (95)	219.0	11.25	<.001	.82
No-Go Accuracy	Points (99)	Non-game (95)	4544.0	0.40	.689	.03

Appendix I

Experiment 1: Does No-Go accuracy get lower as the test goes on? Does this effect differ between task variants?

I calculated mean No-Go accuracy in each sub-block (3 No-Go trials in each), across the task variants. A two-way ANOVA of No-Go accuracy showed clear evidence of a large effect of sub-block number ($F_{24,75}$=3.807, p<.001, η^2=.55) and task variant ($F_{2,75}$=1463.407, p<.001, η^2=.975). There was no evidence for an interaction (p=.071). I used linear regression to investigate the relationship between sub-block number and No-Go accuracy in task variant and found strong evidence for a relationship in each case (non-game: r^2=.16, $F_{1,48}$=9.361, p=.004; points: r^2=.25, $F_{1,48}$=16.175, p<.001; theme: r^2=.33, $F_{1,48}$=23.473, p<.001). I scatterplotted the data and regression lines (Appendix I Figure 1).



Appendix I Figure 1: *Mean sub-block No-Go accuracy plotted against sub-block number. Shown separately by task variant.*

Appendix J

Gamified stop-signal task design decisions

Experiments 2 and 3 both used gamified stop signal tasks (for details of the stop-signal task (SST) see Section 5.3.3). This appendix provides further detail on the game design elements used, and on the design decisions that led to their implementation. In Experiments 2 and 3, as in Experiment 1, I investigated two game design elements: points and theme, and I had five goals when designing these variants:

1. Implement gamification as richly as possible, while not straying from the game design element under investigation
2. Implement gamification which makes the task appear like a game
3. Implement gamification aligned with the task's goals
4. Implement gamification which minimises potential negative effects on cognitive data
5. Implement gamification to encourage an increased amount of engagement

Points Variant

For a demo of the points task variant of the SST see: goo.gl/UmLgVc, a shorted URL which links to mindgamesmkii.firebaseapp.com/task.html?prolific_pid=1873ae62fe8e72e9f868d720

1. Implement gamification as richly as possible, while not straying from the game design element under investigation

In Experiments 1,2 and 3 my intention was to investigate the effects of individual game design elements, however the subjective nature of the term 'game design element' makes this difficult. What does it mean to test the effect of points individually? Can I include gamelike framing? What about a highscore mechanism? Does the aesthetically-pleasing display of a participant's score constitute a separate game design element to the points themselves? As discussed in Section 7.3.1, we do not have precise definitions of game design elements and so I had to decide on the answers to these questions myself.

I decided that shoehorning the most basic implementation of points into an otherwise completely non-gamified task would not be a meaningful test of gamification. Instead, I tried to maintain a balance between implementing only one game design element and supporting its implementation in such a way that it could be effective at creating intrinsic motivation. Theoretically, I thought that points would provide intrinsic motivation by meeting self-determination theory's (SDT's) competency need [154]. I therefore made design decisions that would promote competency while avoiding the introduction of other motivating factors (such as autonomy or relatedness).

In addition to providing constant feedback on participant performance (the most basic form of points), my points task variant contained the following motivational features:

1. Gamelike framing in the menu screens and instructions (Appendix J Figure 1)
2. A display of current score, a display of the participant's current high score (Appendix J Figure 2C)
3. 'Juicy' animations when bonuses were earned or lost (see demo) [221]
4. A leaderboard of the participant's score over past sessions (Experiment 2 only) (Appendix J Figure 2D)
5. A comparison of the participant's current high score against their score from the previous block (Experiment 3 only) (Appendix J Figure 3)

2. <u>Implement gamification to make the task appear like a game</u>

There is some evidence that simply telling participants the task is a game is enough to change their motivations towards it [277]. To avoid any ambiguity, I explicitly referred to the points task variant as a game, and used gameful language throughout (e.g. play, player, score, rules). I retitled the main menu as "Stopmaster", with a subtitle inviting participants to enter a self-competitive frame of mind (Appendix J Figure 1). The instructions were not overly gamified, but I used casual language, bright colours and exclamation marks to reinforce the lighthearted tone.

Appendix J Figure 1 Two screenshots of the points task variant, demonstrating the gamelike framing of the main menu and instructions.

3. <u>Implement gamification aligned with the task's goals</u>

The scoring system used in the points variant of the SST was very similar to that used in Experiment 1, which in turn was based on that used by Miranda and colleagues [49]. The scoring system also incorporates the findings of Guitart-Masip and colleagues [181] who found that subjects were much more successful in learning active (go) choices when rewarded for them, and passive choices (stop) when punished. Accordingly, participants only gained points on the non-stop-trials, and lost some of their *Bonus* when they failed to inhibit on a stop-trial. On each successful non-stop-trial the participant earned points equal to $Bonus \times 0.2 \times (800 - \mathrm{RT})$, and the number of points gained was displayed briefly in the inter-trial interval.

This *Bonus* was a multiplier (x2, x3, x4...), which increased by 1 every 3 trials but decreased by 3 when the participant failed a stop trial.

This scoring system rewarded participants in line with optimal performance on the task. The SST can produce its best estimate of inhibitory performance when the participant balances fast responding against inhibiting responses whenever possible. Likewise, a high score is attained in the points variant by balancing fast responding against inhibiting responses whenever possible.

4. Implement gamification which minimises potential negative effects on cognitive data

The results of Experiment 1 showed that the points variant had no negative impact on cognitive data. I considered it likely that the points variant's visual similarity to the non-game variant played a role in the lack of effect, and accordingly maintained consistent stimuli between the points and non-game variants of the SST in Experiment 2 and 3.

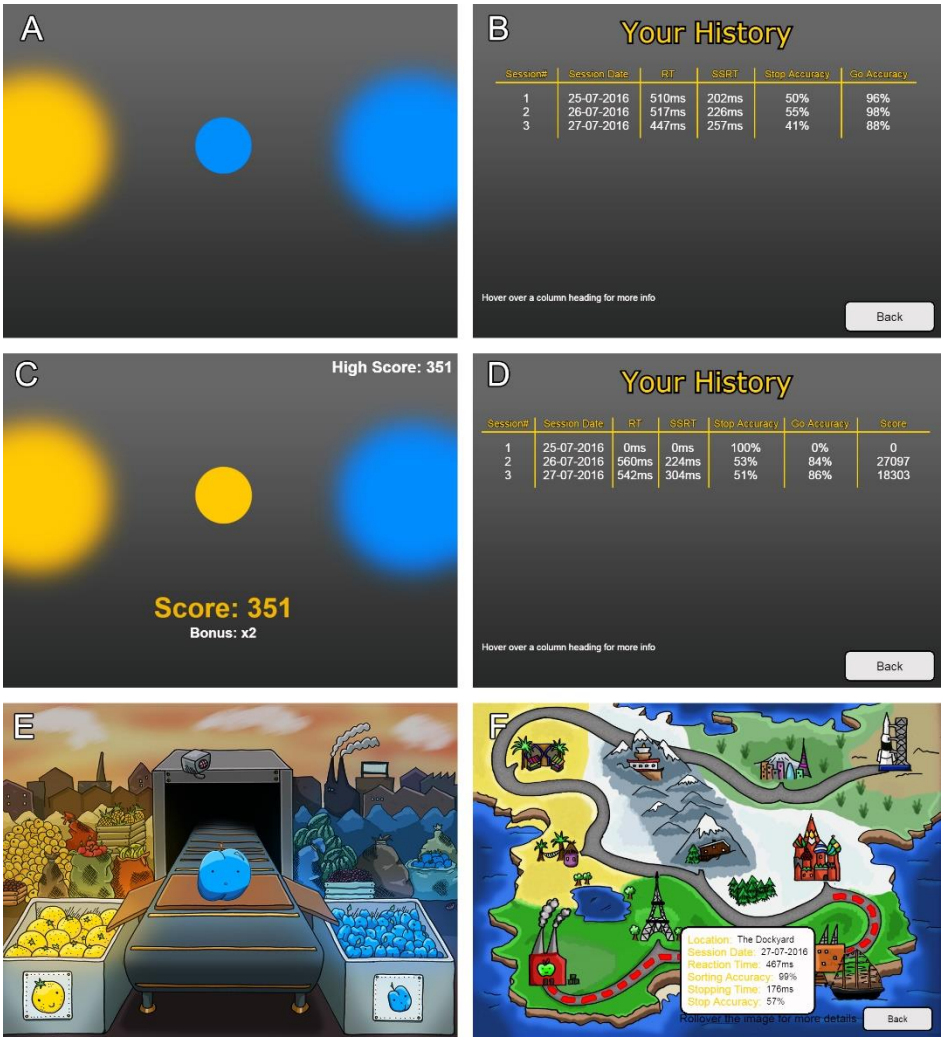
The points variant added only three items to the task screen that were not present in the non-game variant: the participant's current score, bonus and high score (Appendix J, Figure 3). The score was quite large on the screen to make it easier for participants to monitor their performance and to enhance the saliency of the primary game design element. The other two elements were designed to take up minimal screen space.

5. Implement gamification to encourage an increased amount of engagement

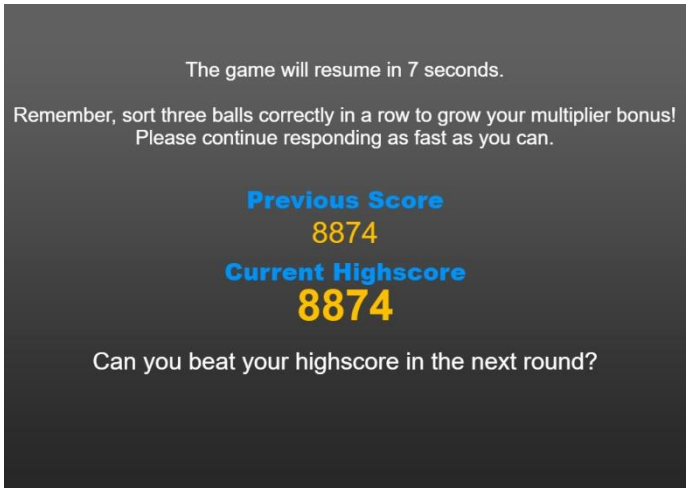
In Experiments 2 and 3 I used points to try and increase amount of engagement. Many gamified applications make use of leaderboards (comparing the user against other users) however I decided not to use this approach for two reasons: 1) leaderboards involve comparison to others, likely inducing motivation through SDT's relatedness-need rather the competency-need which I considered to be the primary motivating factor of the points variant. 2) There is evidence that leaderboards are not always beneficial to participant performance or engagement, with some participants feeling left behind and dropping out, and others aiming to be average rather than striving for the top [155,278].

Instead, I facilitated engagement through self-comparison. The highest score the participant had obtained was displayed in the top right of the screen, as a remainder of a score they could strive for. In Experiment 2, participants could look back over every session they'd completed to compare their scores or seek improvement over time. In Experiment 3, after the participant elected to complete another block of the SST they were explicitly challenged to beat their high score (Appendix J Figure 3). In Experiment 3 I also tried to boost the amount of engagement by slightly increasing the number of points a participant would earn each round: not so much that the participant would automatically beat their high score each round, but enough that they

would be able to beat their high score quite easily after persisting with the SST for a few rounds. There is evidence that accelerating points gain can increase motivation [279].



Appendix J Figure 2 In-task screenshots of the SST variants and the associated history screens from Experiment 2. (A/B) non-game variant, (C/D) points variant, (E/F) theme variant



Appendix J Figure 3 Screenshot of the post-continuation choice screen from the points variant of the SST, challenging the participant to beat their current highscore in the next round

Theme Variant

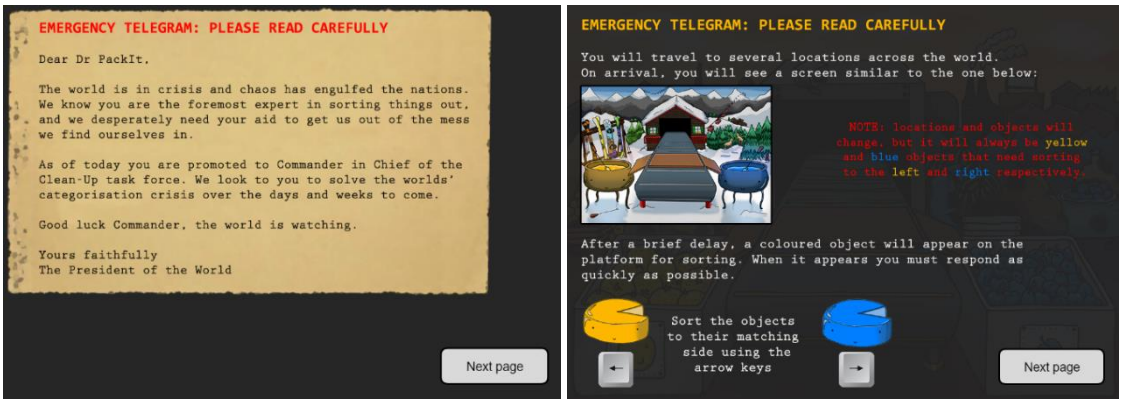
For a demo of the theme task variant of the SST see: goo.gl/ksyacZ, a shorted URL which links to mindgamesmkii.firebaseapp.com/task.html?prolific_pid=7388c04576738aeb16944359

1. ***Implement gamification as richly as possible, while not straying from the intended individual game design element under investigation***

The term 'theme' is commonly used in the gamification literature, but its precise meaning is ambiguous. I considered that the goal of a theme was to provide the participant with a narrative reason as to why they were performing the task, so as to foster a sense of autonomy. To this end, I used a combination of simple graphics and narrative framing to create a cohesive theme. I kept my implementation as simple as possible and consciously avoided complex animation or involved story, to ensure my task remained comparable to other themed cognitive tasks (i.e. [52,97,101]).

2. ***Implement gamification to make the task appear like a game***

As with the points variant, I wanted my gamified theme variant to appear gamelike from the outset. Due to the coloured graphics on the main menu (Figure 5.1) this was easy to achieve, but nevertheless I supported the task's graphics with narratively-framed instructions (Appendix J Figure 4). The participant was immediately placed in the role of Dr PackIt and the SST was explained from an in-game perspective. This was intended to promote immersion in the game-world, thus facilitating engagement.



Appendix J Figure 4 Screenshots of the first and second instructions screens from the theme variant of the SST, showing the narrative framing and whimsical storyline.

3. ***Implement gamification aligned with the task's goals***

In the non-game SST the participant is told they must sort coloured circles to either side of the screen, and that they must occasionally withhold their response to the stop signal, but no reasoning is given for *why* they must perform this task. Gamification wraps the mechanism of the SST in metaphor, providing a narrative reason as to why the participant must perform

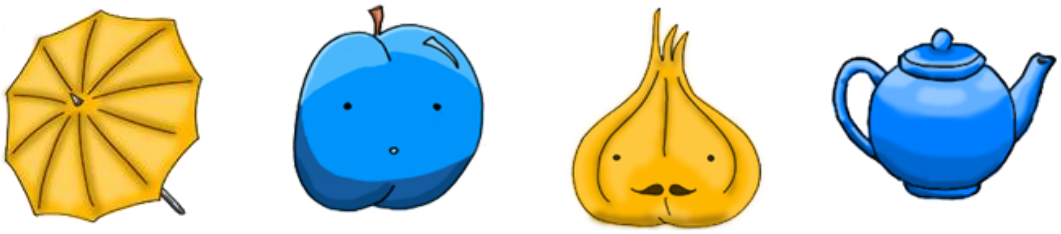
those actions. The narrative reason can be serious, whimsical or outright silly, but the participant is expected to suspend-disbelief and embrace the game-world [244].

In the theme variant, the game opened with a letter from 'the president of the world', imploring the participant's aid in "sorting the world out". A series of light-hearted instructions-screens explained that the world was a mess, and that the participant needed to travel to several global destinations, sorting blue and yellow objects into their respective piles (Appendix J Figure 2E), but not sorting those objects which the 'scanner detected to be faulty'. This narrative framework is an obvious analogy for the actions required by the SST, but it is intuitive and provides a coherent in-game reason of the participant's required actions. Malone described the aligning a gamified theme with the underlying mechanics of the task as 'intrinsic fantasy' [123], and posited that it is an effective way of increasing engagement.

4. <u>Implement gamification which minimises potential negative effects on cognitive data</u>

In Experiment 1, the theme variant showed large negative effects of gamification on RT and no-go accuracy. In retrospect I think this was due to the complexity of the cowboy stimuli, and the difficulty of distinguishing the go stimuli from the no-go stimuli: particularly because colour could not be used to help distinguish the two.

To correct for this in the SST I aimed to keep the theme-variant stimuli as similar as possible to those from the non-game and points variants. I used colour: yellow and blue, to distinguish the left and right stimuli in all three gamified task variants. Furthermore, though the stimuli in the theme variant changed with each location visited, the stimuli's predominant colours were matched between the task variants and their shape was mostly round (Appendix J Figure 5).



Appendix J Figure 5 Selected stimuli from the theme variant. I aimed to minimise negative effects on the cognitive data by matching the colour and shape of themed-stimuli to the non-game and points stimuli.

In addition, though the location background-graphics were brightly coloured and complex, they became somewhat desaturated and darkened once the SST itself began.

5. <u>Implement gamification to encourage an increased amount of engagement</u>

As with the points variant, I intended the theme variant to increase amount of engagement. I attempted this in different ways in Experiments 2 and 3, due to the studies' different timescales (over days and over minutes).

At the end of each session in Experiment 2 I presented the participant with a history screen showing a summary of their performance on previous sessions. In the theme variant this history screen took the form of a map showing their progress through various locations towards their final destination (Appendix J Figure 2F). The purpose of this map was to create an overarching goal which would encourage participants to complete all ten sessions [221]. The map icons hint at upcoming locations with the aim of creating perceptual curiosity and increasing intrinsic motivation [76,123]. Each session of the study featured a new location and a new graphical task-background (Appendix J Figure 7).

On loading each level the participant was shown a brief extract of 'flavour text' which contributed to the overall narrative arc of Dr PackIt sorting out the world (Appendix J Figure 6). I did not attempt to weave a complex or meaningful story so as not to step beyond the bounds of 'theme' as the single game design element under investigation. The flavour text was designed purely to maintain the gamelike atmosphere, and to narratively string the participant between the various sessions of the SST. The whimsical style of graphics and story was chosen to appeal to the broadest range of people. All game design elements were created by me and a contracted artist.

Appendix J Figure 6, Screenshot of the flavour text displayed upon visiting Paris. The text is 'spoken' by the commander's aide, Harper, who introduces each location and reminds the participant of the task's instructions.

The map screen was not used in Experiment 3 because participants played each task variant only once. Instead, I gave participants the chance to see multiple locations in a single test

session: choosing a new location and new task-background at the start of each block (Appendix J Figure 7). By allowing the participant to choose which location they visited, my intention was to provide a sense of autonomy, and accordingly, increase engagement. However, because the participant had control over the route they took through the available locations, I could not easily create a map of their progress. The theme variant in Experiment 3 therefore missed out on the potentially motivating effect of an overarching end goal.

Appendix J Figure 7, Screenshots of all ten task backgrounds in the theme variant. From top left to bottom right: Packville, Paris, The Alps, Hawaii, Moscow, Tokyo, Nepal, International Space Station, Morocco and London.

Appendix K

Experiment 2: Stop Signal Task: Staircase and Block Details

Further to the information provided in Chapter 5: Stop Signal Delay (SSD) was varied according to a four-staircase convergence algorithm, designed to sample evenly across the SSD/Inhibition-Probability space. Staircases 2 and 3 converged to a 50% failed inhibition rate, while staircases 1 and 4 sampled the limits of a participant's inhibition (Appendix K, Table 1). On a step-up or step-down a staircase was adjusted by +/-50 ms respectively, and the step size changed to +/-25ms after two reversals of direction. The shortest possible SSD was 25ms and the longest possible was 750ms.

The task consisted of 5 blocks of 48 trials each. Each block contained 3 sub-blocks of 16 trials each, of which 12 were go trials and 4 were stop trials. The first sub-block of each session consisted entirely of Go trials, so in total each session contained 240 trials, of which 56 were stop trials. After 48 trials the block ended and the subject had to wait for 10 seconds before they automatically continued to the next block. To maintain response speed and to discourage strategy, the subject was prompted to go faster during this break. A dynamic speed-prompt was also displayed if the subject's responses in one sub-block were on average 50 ms slower than those in the previous sub-block. Once five blocks had been completed, the task ended. This typically took ~10 minutes.

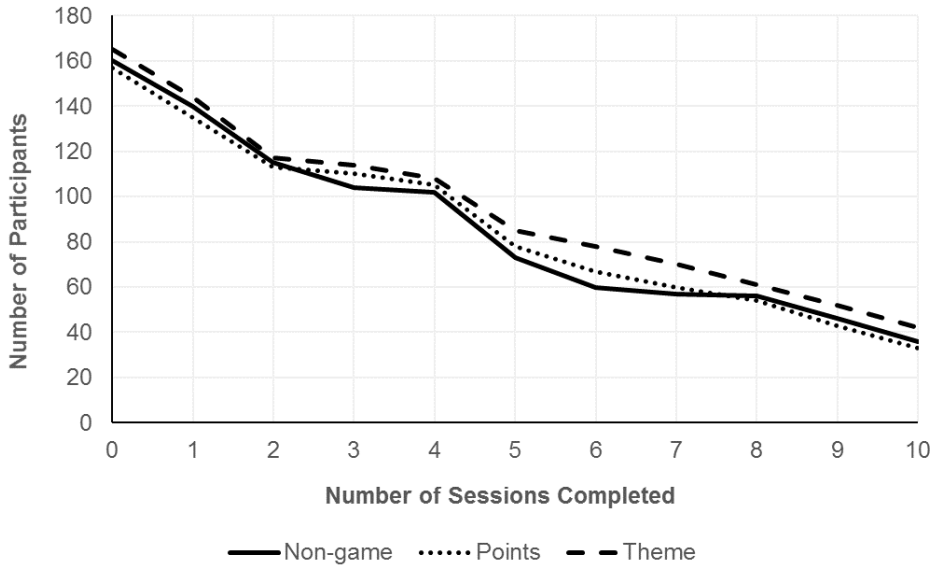
Appendix K Table 1 Stop signal delay staircase initial values. *Due to staircases 1 and 4 tracking the lower and upper limits of inhibition respectively, they require a different number of failed/successful inhibitions to step up or down.*

Staircase number	Initial SSD	Failure rate goal	Number of failed trials in a row needed to step down	Number of successful trials in a row needed to step up
1	150ms	~30%	1	2
2	250ms	~50%	1	1
3	350ms	~50%	1	1
4	400ms	~70%	2	1

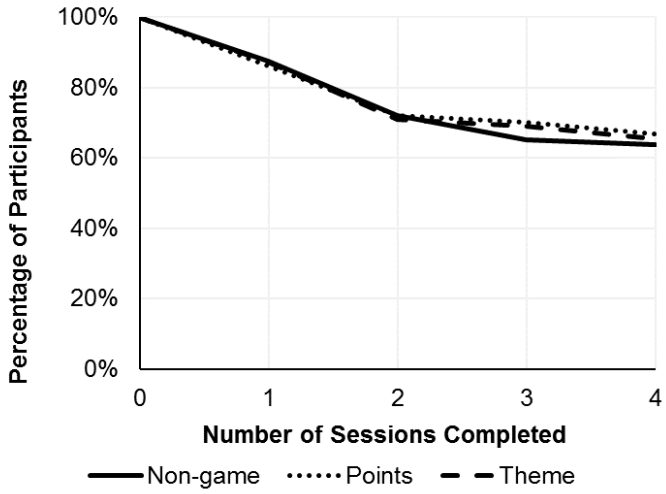
Appendix L

Experiment 2: All Participant Attrition Analysis

Appendix L Figure 1 shows the number of participants remaining in the study at each timepoint, including all participants who signed up. I used the Kaplan Meier method to estimate survival times, and a Log-Rank test showed no evidence of a difference between the distributions ($X^2_{2,482}$=.816, p=.67). I was also interested in whether gamification would affect the number of participants who decided to stay with the study after trying one initial session. Appendix L Figure 2 shows the percentage of participants that completed a session on the first four days, divided by task variant.

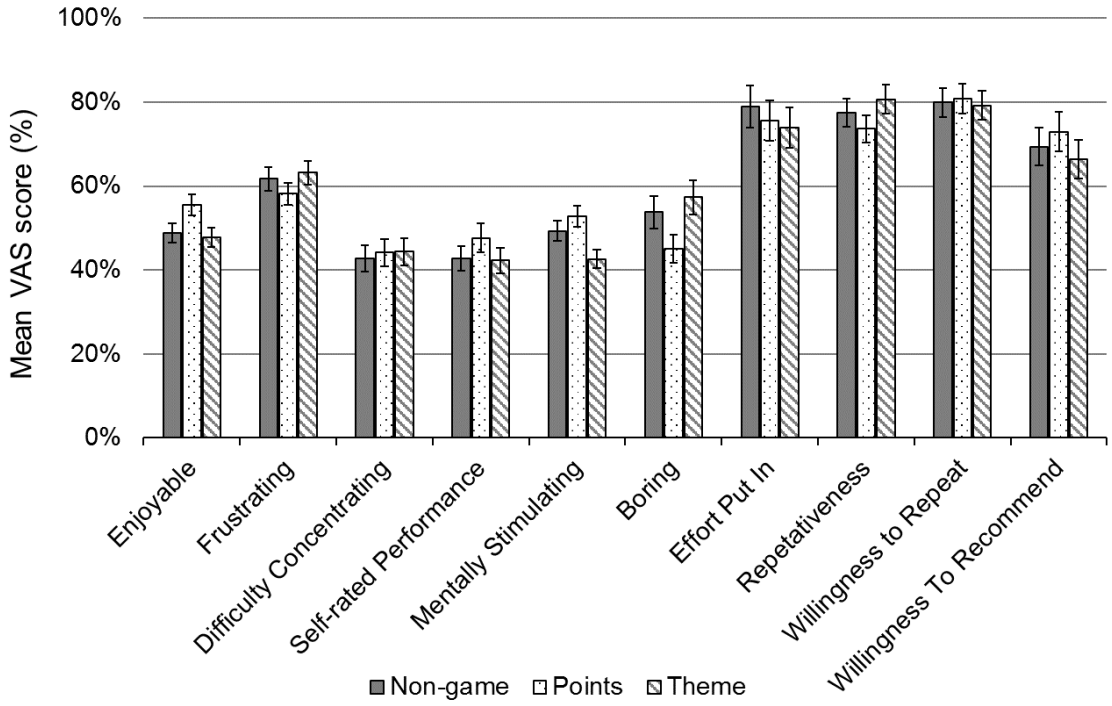


Appendix L, Figure 1*: Number of participants that took part each day over the ten-day period*



Appendix L, Figure 2: Percentage of participants that took part in the study each day over the first four days

Appendix M



Individual question scores from the assessment of quality of engagement. Mean scores calculated from questionnaires delivered on sessions 1 and 4, shown separately by task variant. Error bars represent 95% confidence intervals.

Appendix N

Planned Analyses from Experiment 2

The following analyses were planned as described in my preregistered study protocol: osf.io/ysaqe. These analyses assume a somewhat passing knowledge in the inner workings of the stop-signal task (SST). I recommend Band et al,. and Logan's User Guide [213,220] as great primers on the task.

FailedStop RTs are RTs on trials with a stop signal, but where the participant fails to inhibit their response. They are particularly important in the SST for defining the inhibition curve. According to the race model FailedStop RTs should always be short than the median GoRT, as they represent the 'go process' finishing quickly, before the 'stop process' has time to inhibit it.

Experiment 2: What is the effect of gamification secondary cognitive measures?

Go RTs and FailedStop RTs were summarised at a participant level using medians. I used numerical differentiation to calculate the gradient of the inhibition function at the point where the participant's probability of inhibiting to a stop-signal was 50%.

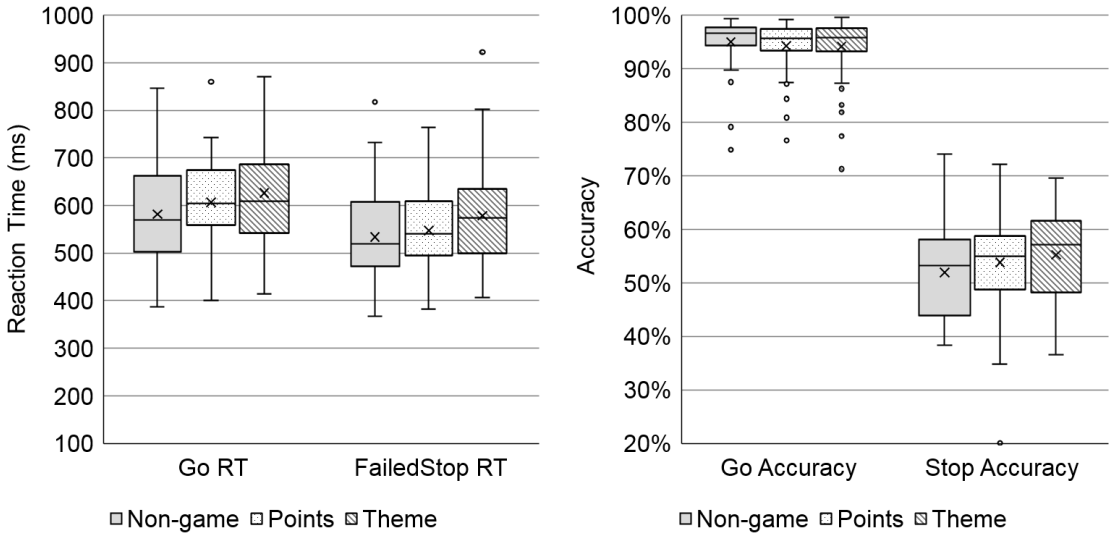
To assess whether gamification affected the cognitive data collected by each task variant I used mean Go RT, FailedStop RT, Go Accuracy and Stop Accuracy data from the four compulsory sessions and performed a series of univariate ANOVAs with task variant (non-game, points, theme) as a factor (Appendix N Table 1). I found clear evidence of an effect of task variant on all measures except for Go Accuracy, and this is likely because Go Accuracy scores were high and participants were performing at ceiling. The effects of task variant were quite small, yet still indicate an impact of gamification on the comparability of the data collected by the task.

Appendix N, Table 1: *Effects of task variant on Go Reaction Time, FailedStop Reaction Time, Go Accuracy and Stop Accuracy. Four univariate ANOVAs on cognitive measures from the first four sessions, with task variant (non-game, points, theme) as a between-subjects factor.*

Dependant Variable	**$F_{2,255}$**	***p***	**partial η^2**
Go RT	4.421	.014	.032
FailedStop RT	5.403	.005	.040
Go Accuracy	1.053	.350	.008
Stop Accuracy	4.450	.013	.033

Appendix N Figure 1 shows boxplots of these variables for each task variant, made of up participants' median responses over the four compulsory sessions. Cognitive measures appear broadly comparable between task variants, but the effects detected by the ANOVAs are

apparent on closer inspection. I used *t*-tests to explore differences of interest, and Bayesian *t*-tests to assess similar distributions for equality.



Appendix N Figure 1: *Box and whisker plots of mean Go Reaction Time, FailedStop Reaction Time, Go Accuracy and Stop Accuracy. Data combined per participant over the first four sessions and shown separately by task variant.*

I found evidence of a medium difference in Go RT between the Non-Game (M=583, SD=100) and Theme (M=622, SD=92) variants (mean difference=38, 95% CI 10 to 67, t_{174}=2.651, *p*=.01, *d*=.41), but no evidence of other differences (*p*s>.12). A Bayesian *t*-test for equality in Go RT distributions between the Non-Game and Point variants was inconclusive (Bayes Factor (BF)=.89).

I also found evidence of a difference in FailedStop RT between the Non-Game (M=530, SD=88) and Theme (M=570, SD=87) variants (mean difference=40, 95% CI 16 to 65, t_{174}=3.068, *p*=.01, *d*=.46), but little evidence of other differences (*p*s>.10). Again, a Bayesian *t*-test could not provide evidence of equality between the non-game and points variants (BF=.42).

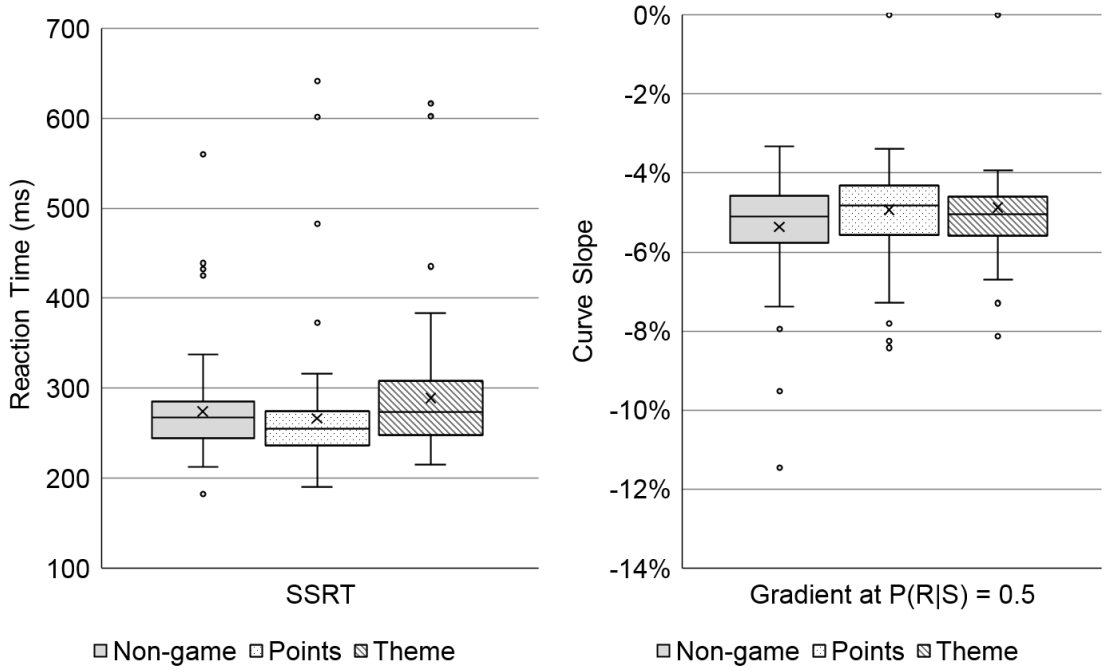
Given the lack of effect of task variant on Go Accuracy I used Bayesian *t*-tests to assess the variants for equality. These tests were inconclusive for all comparisons (BF=.31 and .38) except points (M=94%, SD=5%) compared to theme (M=94%, SD=6%), where I found substantial evidence of equality (BF=.17)

With respect to Stop Accuracy I saw differences between the non-game (M=52%, SD=9%) and theme (M=54%, SD=7%) variants (mean difference 25%, 95% CI 21 to 48, t_{174}=2.153, *p*=.03, *d*=.32) and the non-game and points (M=55%, SD= 8%) variants (mean difference=33%, 95% CI 9 to 57, t_{175}=2.722, *p*=.01, *d*=.40). There was no evidence of a difference between points and theme (*p*=.50), but a Bayesian *t*-test showed no evidence for equality either (BF=.50).

Appendix N, Table 2: *Mean Stop Signal Reaction Times, and inhibition function midpoint-gradients from the first four sessions, shown separately by task variant.*

Task Variant	**SSRT (95% CI)**	**Gradient of the inhibition function at the point of 50% inhibition probability (95% CI)**
Non-Game	274ms (262 to 285)	-5.36% (-5.10 to -5.64)
Points	262ms (250 to 275)	-5.10% (-4.87 to -5.32)
Theme	286ms (273 to 299)	-5.22% (-5.05 to -5.38)

I calculated the slopes of the modelled inhibition curves using numerical differentiation and assessed the gradient for differences between task variants. A one-way ANOVA did not show evidence of an effect of task variant on inhibition slope ($F_{2,255}$=1.437, p=.24, *partial* η^2 =.011), and Bayesian *t*-tests showed moderate evidence that the non-game and theme variants' slopes were equivalent (BF=.24), and that points and theme variants' slopes were also equivalent (BF=.22) However, there was insufficient evidence to suggest that the non-game variant and the points variant had equivalent slopes (BF=.62) (Appendix N Table 2) (Appendix N Figure 2)



Appendix N, Figure 2: *Box and whisker plots of mean Stop Signal Reaction Time and mean Inhibition Function gradient. Data combined per participant over the first four sessions and shown separately by task variant*

Experiment 2: How reliable are cognitive measures over time?

I found the test-retest reliability of stop signal reaction times (SSRTs) from the first four sessions to be very good, with an overall Cronbach's alpha of 0.85. When assessed by task variant, the points (α=.86), and theme (α=.86) variants showed the most consistent results with non-game (α=.75) showing lesser, yet still good, reliability. I used *cocron* [280] to investigate differences between these alphas but saw only weak evidence for an effect of task variant ($X^2_{2,258}$=5.140, p=.08).

I also wanted to investigate whether time or practice effects impacted the cognitive data collected by the task variants, and so ran a series of repeated-measures ANOVAs with Go RT, FailedStop RT and SSRT as the dependant variables and session number (1-4) as the time factor in each (Appendix N Table 3). Where there was evidence of violated sphericity I used Greenhouse-Geisser corrected *p* values. I found small effects of session number on all cognitive measures, but no clear evidence of interactions between task variant and session number on any of the measures (*p*s>.07). Appendix N Table 4 shows the mean RTs from each session, combined across task variant.

Appendix N, Table 3: *Effect of session number on Go Reaction Time, FailedStop Reaction Time and Stop Signal Reaction Time. Three repeated-measures ANOVAs with session number (1-4) as the time-factor and task variant (non-game, points, theme) as the between-subjects factor. Where there was evidence of sphericity I report Greenhouse-Geisser corrected p values.*

Dependant Variable	**$F_{3,762}$**	***p***	**partial η^2**
Go RT	3.336	.025	.013
FailedStop RT	4.822	.004	.019
SSRT	5.139	.003	.033

Appendix N, Table 4: *Mean Go Reaction Times, FailedStop Reaction Times and SSRTs, shown separately by session number.*

Dependant Variable	**Session 1 (95% CI)**	**Session 2 (95% CI)**	**Session 3 (95% CI)**	**Session 4 (95% CI)**
Go RT	601ms (591 to 611)	614ms (602 to 626)	605ms (592 to 618)	602ms (589 to 615)
FailedStop RT	540ms (531 to 549)	556ms (544 to 568)	552ms (540 to 564)	554ms (541 to 567)
SSRT	273ms (266 to 280)	266ms (256 to 276)	259ms (248 to 270)	258ms (248 to 268)

Experiment 2: Do perseverance or individual differences affect attrition?

After the second session of the study participants completed a visual-analogue-scale based perseverance subscale of the Urgency, Premeditation, Perseverance and Sensation Seeking (UPPS) Impulsive Behaviour Scale [281], presented in the same format at the quality of engagement questionnaire. The main aim of this questionnaire was to test whether individual differences in perseverance might confound attrition rates on the task variants. A total perseverance score was calculated as the mean of all items, with items 2 and 10 reverse-scored. The following questions were presented in a random order: (1) I generally like to see things through to the end, (2) I tend to give up easily, (3) Unfinished tasks really bother me. (4) Once I get going on something I hate to stop. (5) I concentrate easily. (6) I finish what I start. (7) I'm pretty good about pacing myself so as to get things done on time. (8) I am a productive person who always gets the job done. (9) Once I start a project, I almost always finish it, and (10) There are so many little jobs that need to be done that I sometimes just ignore them all.

To ensure that individual differences in participant perseverance between groups were not masking an effect of task variant on attrition, I used a one-way ANCOVA of mean number of sessions completed with task variant (non-game, points, theme) as the between-subjects factor and score on the perseverance questionnaire as the covariate. I still saw no evidence of an effect of task variant on the mean number of sessions completed ($F_{2,259}$=1.168, p=.31, *partial* η^2 =.009) and only weak evidence for an effect of perseverance ($F_{1,259}$=3.562, p=.06, *partial* η^2 =.013).

Previous literature has suggested that participant's age, sex or amount of video game experience can impact their enjoyment of a video game, so I ran a one-way ANCOVA of quality of engagement score with task variant (non-game, points, theme) as the between-subjects factor and age, sex and hours spent playing video games as covariates. I found no evidence for any influence of the three covariates (ps>.28) but saw evidence of a small effect of task variant on overall score ($F_{2,259}$=4.030, p=.02, *partial* η^2 =.030).

Appendix O

Analysis of secondary cognitive data for Experiment 3

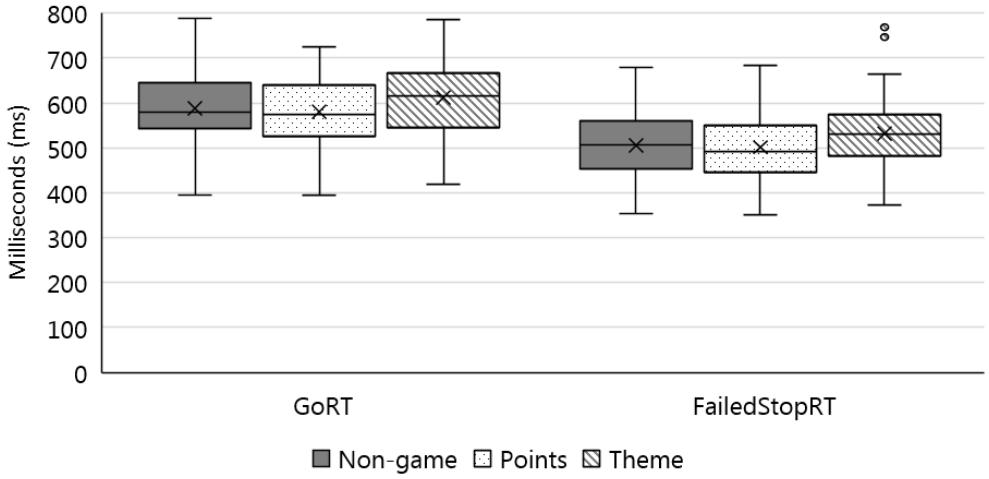
What is the effect of gamification secondary cognitive measures?

Go RTs and FailedStop RTs were summarised at a participant level using medians. To assess whether gamification affected the cognitive data collected by each task variant I used mean Go RT, FailedStop RT and Go Accuracy data and performed a series of univariate RM-ANOVAs with task variant (non-game, points, theme) as a within-subjects factor (Appendix O Table 1). I found clear evidence of large effects of task variant on all measures. Appendix O Figures 1 and 2 show boxplots of these variables for each task variant. The negative effects of theme on cognitive data are apparent, particularly on GoAccuracy where many outliers are present.

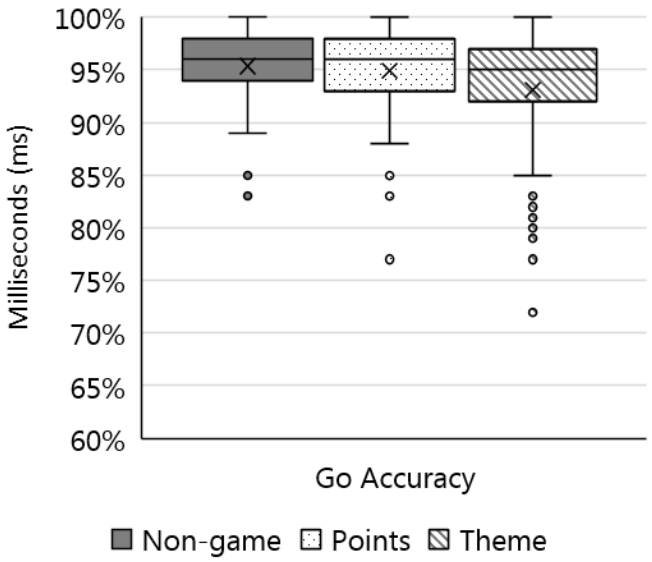
I used Bayesian *t*-tests to weigh the evidence for equality between the non-game and points task variants. I found substantial evidence for equality all measures: GoRT (Bayes Factor (BF)=.20), FailedStopRT (BF=.14), GoAccuracy (BF=.18).

Appendix O, Table 1: *Effects of task variant on Go Reaction Time, FailedStop Reaction Time and Go Accuracy. Four univariate RM-ANOVAs on mean cognitive measures, with task variant (non-game, points, theme) as the within-subjects factor.*

Dependant Variable	**$F_{2,144}$**	***p***	**partial η^2**
Go RT	10.83	<.001	.131
FailedStop RT	8.34	<.001	.104
Go Accuracy	6.94	.001	.088



Appendix O, Figure 1: *Box and whisker plots of mean Go Reaction Time and FailedStop Reaction Time. Data combined over reimbursement schemes and shown separately by task variant.*



Appendix O, Figure 2: *Box and whisker plots of mean Go Accuracy. Data combined over reimbursement schemes and shown separately by task variant.*